

(19)日本国特許庁(J P)

(12) 公 開 特 許 公 報 (A)

(11)特許出願公開番号

特開平6-202817

(43)公開日 平成6年(1994)7月22日

(51)Int.Cl.⁵

G 0 6 F 3/06

識別記号

3 0 5 C 7165-5B

3 0 1 Z 7165-5B

庁内整理番号

F I

技術表示箇所

審査請求 未請求 請求項の数34 (全 33 頁)

(21)出願番号 特願平4-348301

(22)出願日 平成4年(1992)12月28日

(71)出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72)発明者 角田 仁

東京都国分寺市東恋ヶ窪1丁目280番地

株式会社日立製作所中央研究所内

(74)代理人 弁理士 薄田 利率

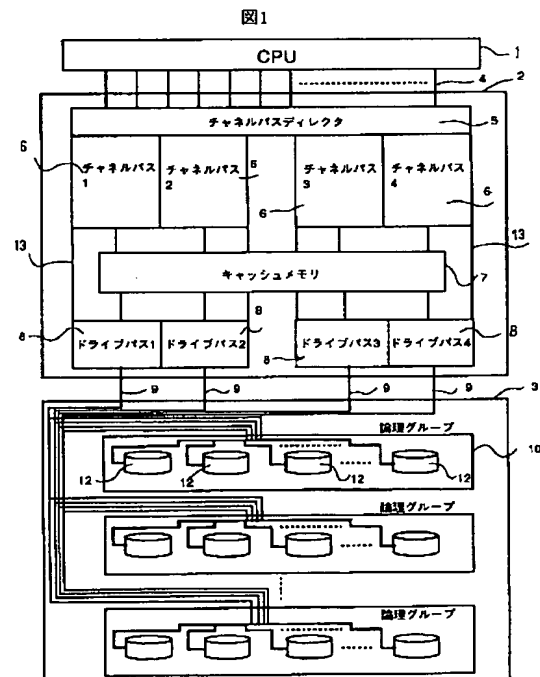
(54)【発明の名称】 ディスクアレイ装置及びそのデータ更新方法

(57)【要約】

【目的】 ディスクアレイシステムにおいて、データの書き込み時のオーバーヘッドを減少させる。

【構成】 パリティグループを構成する論理グループ10に、複数のダミーデータ領域を含ませる。1つの論理グループの各データ及びパリティは、各々別々のSCSIドライブ12に分散して格納される。更新時において、更新するデータ及びパリティを書き込む際に、最も効率良く書き込めるSCSIドライブを判定し、スケジューリングを行なう。更新するデータ及びパリティは更新前のデータ、パリティが格納されていた領域ではなく、このダミーデータの位置に書き込む。

【効果】 データの更新時において、更新するデータ及びパリティの書き込みをスケジューリングしてダミーデータ領域に効率良く書き込むため、書き込み時のオーバーヘッドは減少する。また、障害発生時のデータ回復処理や、ドライブ拡張時のデータ書き込み処理が容易となる。



【特許請求の範囲】

【請求項1】それぞれ複数のデータと、該複数のデータから生成された少なくとも1つの誤り訂正符号を含む複数のデータグループを生成し、各データグループを構成する複数のデータ及び誤り訂正符号が複数のドライブ内に分散されるようにして保持する記憶装置におけるデータの更新方法であって、

いずれかのデータグループを構成する複数のデータの1つを新たなデータで書き換えるときに、その一つのデータを新たなデータで更新した後の、該データグループに対する新たな誤り訂正符号を新たに生成し、該新データと該新誤り訂正符号とを、前記記憶装置に含まれた各データグループを分散して書き込むに必要なドライブの数より多いドライブの各々に設けられた予備領域の内、該更新すべきデータグループに対するデータと誤り訂正符号がそれぞれ記憶されているドライブとは異なるドライブに、分散して保持することを特徴とする記憶装置におけるデータの更新方法。

【請求項2】前記書きかえられるべきデータと、該データの属するデータグループの旧誤り訂正符号とを保持していたドライブの領域を新たな予備領域として確保することを特徴とする請求項1記載の記憶装置におけるデータの更新方法。

【請求項3】それぞれ複数のデータと、該複数のデータから生成された少なくとも1つの誤り訂正符号を含む複数のデータグループを生成し、各データグループを構成する複数のデータ及び誤り訂正符号が複数のドライブ内に分散されるようにして保持する記憶装置におけるデータの更新方法であって、

それぞれのデータグループの書き込み時に、それぞれのデータグループに対して、該データグループ内の各データと誤り訂正符号を書き込むドライブと異なるドライブに、少なくとも2個の予備領域を確保し、

いずれかのデータグループに属するデータのの一つを新たなデータで更新するときに、該データグループに対して確保された前記少なくとも二つの予備領域の一つに該新たなデータを書き込み、

該新たなデータで更新された後の該データグループに対する新たな誤り訂正符号を生成し、

前記新たな誤り訂正符号を前記少なくとも二つの予備領域の他の1つに書き込み、

前記書き換え前のデータ及び誤り訂正符号がそれぞれ書き込まれていた二つの記憶領域を、上記データグループのための新たな予備領域として保持する、

ことを特徴とする記憶装置におけるデータの更新方法。

【請求項4】前記データグループ内において、前記誤り訂正符号を生成する各データは、別々のドライブの同一の物理的なアドレスに格納し、前記誤り訂正符号も同様に別のドライブの、該誤り訂正符号の生成に関与したデータと同一の物理的なアドレスに格納し、

前記予備領域は前記データおよび誤り訂正符号と同様に、前記データ及び誤り訂正符号が格納されているドライブとは別のドライブの同一の物理的なアドレスに各々格納することを特徴とする請求項1または3記載の記憶装置におけるデータの更新方法。

【請求項5】データを格納する複数のドライブ群を有するディスク装置と、外部からのデータの入出力要求に対応して該ディスク装置のデータ格納を管理する制御装置からなるディスクアレイシステムにおけるデータの更新方法であって、該格納されているデータから複数のデータを選択してパリティを生成し、該パリティとその生成に関与したデータの集合をパリティグループの単位とし、各パリティグループのパリティとデータを前記ドライブ群に格納するものにおいて、外部から1回に読み出したまたは書き込みする単位で転送されてきたデータを前記ドライブ群の各格納領域に分散して格納し、

該分散して格納されているデータから複数のデータを選択してパリティを生成し、該パリティとその生成に関与したデータの集合をパリティグループの単位とし、前記各パリティグループに複数のダミーデータを付与し、前記各パリティグループにおける、前記複数のダミーデータを、前記パリティとその生成に関与するデータとが格納されているドライブ群とは別のドライブ群に、各々独立して保持することを特徴とするディスクアレイシステムにおけるデータ更新方法。

【請求項6】データを格納する複数のドライブ群を備えたディスク装置と、上位装置からのデータの入出力要求に対応して当該ディスク装置を管理する制御装置からなるディスクアレイシステムにおけるデータ更新方法において、

上記上位装置から1回に読み出したまたは書き込みする単位で転送されてきたデータを複数のドライブ群に別々に分散して格納し、

該別々のドライブに格納されているデータから複数のデータを選択し、これらの選択されたデータからパリティを生成し、

この生成したパリティと生成に関与したデータの集合をパリティグループとし、同一のパリティグループにおいては、パリティもデータと同様に、このパリティを生成するのに関与したデータが格納されているのとは別々のドライブに各々格納すると共に、

各パリティグループに対する複数の予備の書き込み領域を、パリティを生成するデータと、それらのデータにより生成したパリティが格納されているドライブに各々分散して保持することを特徴とするディスクアレイシステムにおけるデータ更新方法。

【請求項7】予備の書き込み領域を1パリティグループに対し少なくとも2個以上対応させ、この予備の書き込み領域において、特定のパリティグループに対応した

予備の書き込み領域をダミーデータとし、前記パリティグループにおいてパリティの生成に関与するデータと、それらのデータより生成したパリティとが格納されているドライブとは別のドライブに、前記それぞれのダミーデータを格納することを特徴とする請求項5または6記載のディスクアレイシステムにおけるデータ更新方法。

【請求項8】パリティを生成するデータと生成したパリティにより構成されるパリティグループと、予備の書き込み領域が、それぞれ別のドライブに格納されており、これらのドライブの集合を論理グループとし、該論理グループ内において、パリティを生成する各データは、別々のドライブの同一の物理的なアドレスに格納し、パリティも同様に別のドライブの、パリティの生成に関与したデータと同一の物理的なアドレスに格納し、ダミーデータはデータおよびパリティと同様に、データ、パリティが格納されているドライブとは別のドライブの同一の物理的なアドレスにもつことを特徴とする請求項5または6記載のディスクアレイシステムにおけるデータ更新方法。

【請求項9】パリティを生成するデータと生成したパリティにより構成されるパリティグループと、予備の書き込み領域が、それぞれ別のドライブに格納されており、これらのドライブの集合を論理グループとし、前記ディスク装置を制御する制御装置は、上位装置からの書き込み要求に対し、論理グループ内において、上位装置が書き込み先として指定してきたデータのアドレスを前記ダミーデータの物理的なアドレスに変換して書き込むことを特徴とする請求項5または6記載のディスクアレイシステムにおけるデータ更新方法。

【請求項10】パリティを生成するデータと生成したパリティにより構成されるパリティグループと、予備の書き込み領域が、それぞれ別のドライブに格納されており、これらのドライブの集合を論理グループとし、前記ディスク装置を制御する制御装置は、上位装置から書き込み要求が発行された場合、上位装置が書き込み先に指定したアドレスに対応する物理的なアドレスに書き込まれている旧データが格納されているドライブが属する論理グループにおいて、その論理グループ内の旧データおよび更新される旧パリティおよびダミーデータが格納されている各ドライブに対して、各々の回転位置を検出し、該検出された回転位置の情報から、回転待ち時間を算出し、該回転位置または回転待ち時間の情報から、ダミーデータの格納されている複数のドライブ群の中で、どのドライブ内のダミーデータの物理的なアドレスに論理グループの新たなデータを格納するかを判定することを特徴とする請求項5または6記載のディスクアレイシステムにおけるデータ更新方法。

【請求項11】前記ディスク装置がキャッシュメモリを備えており、パリティを生成するデータと生成したパリティにより構成されるパリティグループと予備の書き込み領域がそれぞれ別のドライブに格納されており、これらのドライブの集合を論理グループとし、前記論理グループにおいてダミーデータの格納されている複数のドライブ群の中から、実際に新データを格納するドライブを判定した後、そのドライブに対し先行して書き込み要求を発行して新データを書き込み、同時に前記キャッシュメモリ内に新データのコピーを残し、該キャッシュメモリ内の新データのコピーにより、新パリティの作成を行い、該新パリティの作成完了後、この新パリティを新データを格納したとは別のダミーデータの物理的なアドレスに書き込むことを特徴とする請求項5または6記載のディスクアレイシステムにおけるデータ更新方法。

【請求項12】パリティを生成するデータと生成したパリティにより構成されるパリティグループと、予備の書き込み領域が、それぞれ別のドライブに格納されており、これらのドライブの集合を論理グループとし、前記論理グループ内のあるドライブに障害が発生したとき、該障害が発生したドライブ内のデータまたはパリティを回復処理により復元し、この復元したデータまたはパリティを予備の書き込み領域に書き込み、前記障害が発生したドライブを正常なドライブに交換し、前記交換後、この交換した正常なドライブは全てダミーデータにより構成される予備の書き込み領域により構成されているとして論理グループを再構成して処理を再開することを特徴とする請求項5または6記載のディスクアレイシステムにおけるデータ更新方法。

【請求項13】パリティを生成するデータと生成したパリティにより構成されるパリティグループと、予備の書き込み領域が、それぞれ別のドライブに格納されており、これらのドライブの集合を論理グループとし、前記論理グループ内のあるドライブに障害が発生したことを感知した場合、障害が発生したドライブを正常なドライブに交換し、前記障害が発生したドライブ内のデータまたはパリティを回復処理により復元し、この復元したデータまたはパリティと、障害が発生したドライブ内にあったダミーデータを、交換した正常なドライブに書き込んで論理グループを再構成して処理を再開することを特徴とする請求項5または6記載のディスクアレイシステムにおけるデータ更新方法。

【請求項14】パリティを生成するデータと生成したパリティにより構成されるパリティグループと、予備の書き込み領域が、それぞれ別のドライブに格納されており、これらのドライブの集合を論理グループとし、前記論理グループ内のダミーデータの数を、書き込み処理

10

20

30

40

50

時間の要求にあわせて変更可能としたことを特徴とする請求項5または6記載のディスクアレイシステムにおけるデータ更新方法。

【請求項15】パリティを生成するデータと生成したパリティにより構成されるパリティグループと、予備の書き込み領域が、それぞれ別のドライブに格納されており、これらのドライブの集合を論理グループとし、前記論理グループ内のダミーデータの数、障害発生時に対する信頼性の要求にあわせて変更可能としたことを特徴とする請求項5または6記載のディスクアレイシステムにおけるデータ更新方法。

【請求項16】同一パリティグループに対して、ダミーデータ数の異なるグループを混在させることを特徴とする請求項5または6記載のディスクアレイシステムにおけるデータ更新方法。

【請求項17】前記パリティグループにおけるデータとパリティとダミーデータの構成を、テーブルにより管理することを特徴とする請求項5または6記載のディスクアレイシステムにおけるデータ更新方法。

【請求項18】複数パリティグループ間で、予備領域を共有することを特徴とする請求項5記載の記憶装置におけるデータ更新方法。

【請求項19】パリティを作成するデータの集合と作成されたパリティにより構成されるパリティグループと、ダミーデータでサブ論理グループを構成し、各論理グループをこのサブ論理グループにより構成し、前記複数の論理グループ間でダミーデータを共有する場合、サブ論理グループのデータとパリティとダミーデータの割り当て方を変えることにより実現することを特徴とする請求項5または6記載のディスクアレイシステムにおけるデータ更新方法。

【請求項20】前記サブ論理グループを構成するデータとパリティとダミーデータの割り当てを、テーブルにより管理することを特徴とする請求項19記載のディスクアレイシステムにおけるデータ更新方法。

【請求項21】前記サブ論理グループを構成するデータとパリティとダミーデータの割り当てを、ユーザが自由に設定することを可能とすることを特徴とする請求項19記載のディスクアレイシステムにおけるデータ更新方法。

【請求項22】前記サブ論理グループを構成するデータとパリティとダミーデータの割り当てを、上位装置からの読み出し要求または書き込み要求が、競合しないような論理グループにおいて構成することを特徴とする請求項19記載のディスクアレイシステムにおけるデータ更新方法。

【請求項23】書き込むべき複数のデータから少なくとも1つの誤り訂正符号を含む1つのデータグループを生成し、該データグループを構成する複数のデータ及び誤り訂正符号を複数のドライブ群内に格納する制御装置を有

する記憶装置であって、該制御装置が、前記データグループを構成する複数のデータの1つを新たなデータに書き換える更新要求に応答して、該更新データと新たな誤り訂正符号とを含む1つのデータグループを新たに生成する手段と、

前記データグループの更新データと誤り訂正符号とを、前記書き換え前のデータおよび誤り訂正符号が格納されていたドライブとは異なるドライブにそれぞれ格納する手段とを備えたことを特徴とする記憶装置。

10 【請求項24】書き込むべき複数のデータから少なくとも1つの誤り訂正符号を含む1つのデータグループを生成し、該データグループを構成する複数のデータ及び誤り訂正符号を複数のドライブ群内に格納する制御装置を有する記憶装置であって、該制御装置が、データの書き込み要求に応答して、前記データグループに対して、該データグループ内の各データと誤り訂正符号及び少なくとも2個の予備領域からなる複数の記憶領域を前記記憶装置の互いに異なるドライブ内に各々確保する手段と、

20 前記確保された複数の記憶領域に前記データグループ内の各データと誤り訂正符号を各々書き込み、残りの少なくとも二つの記憶領域を前記予備領域として保持する手段と、

記憶された前記データの一つを新たなデータで更新する要求に応答して、前記少なくとも二つの予備領域の一つに該新たなデータを書き込む手段と、

前記新たなデータに更新された後の複数のデータに対する新たな誤り訂正符号を含む新たなデータグループを生成する手段と、

30 前記新たな誤り訂正符号を前記少なくとも二つの予備領域の他の1つに書き込み、前記複数の記憶領域のうち、書き換えられた前記複数のデータ及び誤り訂正符号が書き込まれていた少なくとも二つの記憶領域を上記更新されたデータグループのための新たな予備領域として保持する手段、

とを備えたことを特徴とする記憶装置。

【請求項25】データを格納する複数のドライブ群を有するディスク装置と、外部からのデータの入出力要求に対応して該ディスク装置のデータ格納を管理する制御装置からなり、該制御装置が、前記格納されているデータから複数のデータを選択してパリティを生成し、該パリティとその生成に関与したデータの集合をパリティグループの単位とし、各パリティグループのパリティとデータを前記ドライブ群に格納するディスクアレイシステムにおいて、

前記制御装置が、

外部から1回に読み出しまたは書き込みする単位で転送されてきたデータを前記ドライブ群の各格納領域に分散して格納する手段と、

50 該分散して格納されているデータから複数のデータを選

択してパリティを生成し、該パリティとその生成に関与したデータの集合をパリティグループの単位とし、前記各パリティグループに複数のダミーデータを付与する手段と、

前記各パリティグループにおける、前記複数のダミーデータを、前記パリティとその生成に関与するデータとが格納されているドライブ群とは別のドライブ群に、各々独立して格納する手段とを備えたことを特徴とするディスクアレイシステム。

【請求項 26】データを格納する複数のドライブ群を備えたディスク装置と、上位装置からのデータの入出力要求に対応して当該ディスク装置を管理する制御装置からなるディスクアレイシステムにおいて、前記制御装置が、

上記上位装置から 1 回に読み出したまたは書き込みする単位で転送されてきたデータを複数のドライブ群に別々に分散して格納する手段と、

該別々のドライブに格納されているデータから複数のデータを選択し、これらの選択されたデータからパリティを生成するパリティ生成回路と、

この生成したパリティと生成に関与したデータの集合をパリティグループとし、同一のパリティグループにおいては、パリティもデータと同様に、このパリティを生成するのに関与したデータが格納されているのとは別々のドライブに各々格納すると共に、各パリティグループに対する複数の予備の書き込み領域を、パリティを生成するデータと、それらのデータにより生成したパリティが格納されているドライブに各々分散して確保する手段とを備えたことを特徴とするディスクアレイシステム。

【請求項 27】前記制御装置は、上位装置からの書き込み要求に対し、パリティグループ内において、上位装置が書き込み先として指定してきたデータのアドレスを前記ダミーデータの物理的なアドレスに変換して書き込むアドレス変換用のテーブルを備えていることを特徴とする請求項 25 または 26 記載のディスクアレイシステム。

【請求項 28】前記制御装置は、上位装置から書き込み要求が発行された場合、上位装置が書き込み先に指定したアドレスに対応する物理的なアドレスに書き込まれている当該旧データが格納されているドライブが属する論理グループにおいて、その論理グループ内の当該旧データおよび更新される当該旧パリティ及びダミーデータが格納されている各当該ドライブに対して、各々の回転位置を検出する回転位置検出回路と、

該検出された回転位置の情報から、回転待ち時間を算出する手段と、

該回転位置または回転待ち時間の情報から、ダミーデータの格納されている複数のドライブ群の中で、どのドライブ内のダミーデータの物理的なアドレスに論理グループの新たなデータを格納するかを判定する手段とを備えていることを特徴とする請求項 25 または 26 記載のデ

ィスクアレイシステム。

【請求項 29】前記制御装置は、上位装置から書き込み要求が発行された場合、パリティグループにおいて当該旧データと当該旧パリティの格納されているドライブ群の中で、回転待ち時間の大きい方のドライブを基準にし、その、基準のドライブから当該旧データまたは当該旧パリティを読み出して、更新する新しいパリティを作成すると共に、ダミーデータの格納されている複数のドライブ群の中で、一番早くその新パリティを書き込めるドライブを判定する手段とを備えていることを特徴とする請求項 25 または 26 記載のディスクアレイシステム。

【請求項 30】前記制御装置は、前記パリティグループのドライブが属する論理グループ内のあるドライブに障害が発生したとき、該障害が発生したドライブ内のデータまたはパリティを回復処理により復元する手段と、該復元データまたはパリティを前記予備の書き込み領域に書き込む手段と、

前記障害が発生したドライブを正常なドライブに交換し、該交換した正常なドライブ群は全てダミーデータにより構成される予備の書き込み領域により構成されているとして論理グループを再構成して処理を再開する手段とを備えていることを特徴とする請求項 25 または 26 記載のディスクアレイシステム。

【請求項 31】前記制御装置は、前記パリティグループのドライブが属する論理グループ内において、データとパリティとダミーデータの構成を管理するテーブルを備えていることを特徴とする請求項 25 または 26 記載のディスクアレイシステム。

【請求項 32】前記制御装置は、サブ論理グループを構成するデータとパリティとダミーデータの割り当てを管理するテーブルを備えていることを特徴とする請求項 31 記載のディスクアレイシステム。

【請求項 33】前記制御装置は、前記パリティグループのドライブが属する論理グループ内のあるドライブに障害が発生した場合、正常な残りのドライブ内のデータとパリティから、障害が発生したドライブ内のデータまたはパリティを回復処理により復元すると共に、この復元したデータまたはパリティを予備の書き込み領域に書き込む制御を行うプロセッサを持つことを特徴とする請求項 25 または 26 記載のディスクアレイシステム。

【請求項 34】前記制御装置は、前記パリティグループのドライブが属する論理グループ単位に、その内部にアドレス変換用のテーブルと、回転位置検出回路と、パリティ生成回路と、キャッシュメモリと、それらを制御するマイクロプロセッサを持つことを特徴とする請求項 25 または 26 記載のディスクアレイシステム。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明はディスクアレイ装置に関するものである。

【0002】

【従来の技術】近年高度情報化に伴い、コンピュータシステムにおいて、2次記憶装置の高性能化が要求されてきた。その一つの解として、多数の比較的容量の小さなディスク装置（以下ドライブ）により構成されるディスクアレイが考えられている。例えば、次の論文にそのような例が開示されている。

【0003】「D. Patterson, G. Gibson, and R. H. Kartz; A Case for Redundant Arrays of Inexpensive Disks(RAID), in ACM SIGMOD Conference, Chicago, I L, (June1988)」

上記論文において、データを分割して並列に処理を行うディスクアレイ（レベル3）とデータを分散して、独立に扱うディスクアレイ（レベル5）について、その性能および信頼性の検討結果が報告されている。

【0004】以下にデータを分散して、独立に扱うディスクアレイ（レベル5）について説明する。レベル5のディスクアレイでは個々のデータを分割せずに独立に扱い、多数の比較的容量の小さなドライブに分散して格納するものである。現在、一般に使用されている汎用大型コンピュータシステムの2次記憶装置では、1ドライブ当りの容量が大きいため、他の読み出し／書き込み要求に当該ドライブが使用されて、そのドライブを使用できずに待たされることが多く発生した。このタイプのディスクアレイでは汎用大型コンピュータシステムの2次記憶装置で使用されている大容量のドライブを、多数の比較的容量の小さなドライブで構成し、データを分散して格納してあるため、読み出し／書き込み要求が増加してもディスクアレイの複数のドライブで分散して処理するため、読み出し／書き込み要求がまたされることが減少する。しかし、ディスクアレイは、このように多数のドライブにより構成されるため、部品点数が増加し障害が発生する確率が高くなる。そこで、信頼性の向上を図る目的で、パリティが使用されている。

【0005】図21は前記論文において D. Patterson らが提案した RAID に述べられている、データを分散して、独立に扱うディスクアレイ（レベル5）内部のデータアドレスを示している。この各アドレスにあるデータは、1回の読み出し／書き込みで処理される単位であり、個々のデータは独立している。また、RAID に述べられているアーキテクチャでは、データに対するアドレスは固定されている。前述したようにこのようなシステムでは、信頼性を向上するためパリティを設定することが不可欠である。本システムでは、各ドライブ内の同一アドレスのデータによりパリティが作成される。すなわち、ドライブ#1から4までのアドレス（1, 1）のデータによりパリティが作成され、パリティを格納するドライブの（1, 1）に格納される。本システムでは読み出し／書き込み処理は、現在の汎用大型計算機システムと同様に、各ドライブに対し当該データをアクセスす

る。すなわち、図22に示すように、当該データが格納されているトラックが所属するシリンダの位置とそのシリンダ内において当該データが格納されているトラックを決定するヘッドアドレスHHと、そのトラック内のレコードの位置を特定する。具体的には要求データが格納されている当該ドライブ12の番号（ドライブ番号）と当該ドライブ内のシリンダ番号であるシリンダアドレス（CC）とシリンダにおいてトラックを選択するヘッド140の番号であるヘッドアドレス（HH）とレコードアドレス（R）からなるCCHHRである。

【0006】このようなディスクアレイにおいて、例えば図21のドライブ#3のアドレス（2, 2）のデータを更新する場合、まず、更新される前のドライブ#3の（2, 2）のデータとパリティを格納してあるドライブの（2, 2）のパリティを読み出し（ステップ1）、これらと更新する新しいデータとで排他的論理和をとり、新たなパリティを作成する（ステップ2）。パリティの作成完了後、更新する新しいデータをドライブ#3の（2, 2）に、新パリティをパリティを格納するドライブの（2, 2）に格納する（ステップ3）。

【0007】このように、ディスクアレイではデータからパリティを作成しデータと同様にドライブに格納しておく。この時、パリティは、パリティの作成に関与したデータとは別のドライブに格納される。この結果、データを格納したドライブに障害が発生した場合、その障害ドライブ内のデータを復元することが可能となる。

【0008】これらのディスクアレイでは、現在一般に使用されている汎用大型コンピュータシステムと同様、2次記憶装置内では、個々のデータの格納場所（アドレス）は予め指定したアドレスに固定され、CPUから当該データへ読み出しまたは書き込みする場合、この固定されたアドレスへアクセスすることになっている。

【0009】このようなレベル5のディスクアレイでは、図23に示すように、データの格納されているドライブ、パリティの格納されているドライブから古いデータとパリティを読み出すため、ディスクを平均1/2回転待ち、それから読みだしてパリティを作成する。この新しく作成したパリティを書き込むため更に一回転必要となり、データを書き替える場合最低で1.5回転待たなければならない。ドライブにおいては1.5回転ディスクの回転を待つということは非常に大きなオーバーヘッドとなる。

【0010】このような書き込み時のオーバーヘッドを削減するため、更新後のパリティグループの書き込み先のアドレスを動的に変換する方法がSTK社から出願されている国際出願公開公報WO 91/20076に開示されている。

【0011】一方、特開平4-230512号公報には、DASDアレイのための更新記録方法及び装置が開示されている。この方法によれば、各パリティ・グルー

ブのデータ・ブロックとパリティ・ブロックが、障害について関連性のない形式でDASDアレイに分散され、かつ各ドライブに複数の論理グループに共通に、未使用スペースが予約される。第1サイクルの間、古いデータ・ブロック及び古いパリティ・ブロックが読み出される。これらのデータと更新用の新しいデータから新しいパリティが計算され、古いデータや古いパリティが保持されていた未使用スペースに、その新しいデータ及び新しいパリティ・ブロックがそれぞれシャドー・ライトされる。

【0012】

【発明が解決しようとする課題】前記国際出願公開報WO 91/20076に開示された動的アドレス変換を行なう場合以下のような問題が生じる。

(1) 動的アドレス変換を行なうと、更新前のパリティグループを保持していた領域がそのままドライブ内に無駄な領域として残され、動的にアドレス変換を行なおうとしても、新たな格納領域を確保できなくなる危険性がある。

(2) アドレス情報をテーブル上に管理するとテーブルの容量が増大しメモリのコストが増加し、また、アドレス制御が複雑となるため、処理オーバーヘッドが大きくなる。

【0013】一方、前記特開平4-230512号公報に記載された未使用スペースを用いる方法によれば、上記(1)の問題はないと考えられる。しかし、上記

(2)のアドレス制御に関しては、その方法が明瞭には開示されていない。さらに、増設時や障害発生時にどのように対処するかについても言及されていない。

【0014】本発明の目的は、動的アドレス変換を行う場合に、新たな格納領域の確保が容易な、データ書き込み方式を提供することにある。

【0015】本発明の他の目的は、動的アドレス変換が可能で、かつアドレス制御が簡単であり、増設時や障害回復時にも柔軟に対処できる、データ書き込み方式を提供することにある。

【0016】

【課題を解決するための手段】本発明によれば、複数のデータと少なくとも1つの誤り訂正コード(例えばパリティコード)を含むパリティグループを構成するドライブ群と、複数の予備の書き込み領域のデータ(ダミーデータ)のドライブ群とにより、1つの論理グループを構成する。この論理グループを構成するドライブ群は、複数のデータとパリティとダミーデータとを各々独立して格納するように構成されている。本発明では、書き込み前と書き込み後では論理グループを構成するドライブ群が異なる点が特徴の1つである。

【0017】すなわち、書き込み後のデータを書き込むドライブと、書き込み後の新たなパリティを書き込むドライブとを、複数のダミーデータを保有するドライブの

中から、書き込み処理全体を見て最も回転待ち時間が少なく、効率的に処理できるように選定する。

【0018】新しいパリティと新しいデータを格納するドライブを決定した後は、新しいデータを格納するドライブへの書き込みが可能になり次第、新しいデータを書込み、旧ドライブ群から書き込み前のデータとパリティを読み出し、書き込む新しいデータとで書き込み後の新しいパリティを作成し、決定済のドライブに書き込み後の新しいデータと新しいパリティを格納する。新しいデータと新しいパリティを格納した後、更新される前のデータが格納されていたドライブと旧パリティが格納されていたドライブとが、各々新しくダミーデータの格納されるドライブとなる。

【0019】本発明の他の特徴によれば、各論理グループにおいて回転位置情報または回転待ち時間の情報に基づいて書き込みデータ(新データ)を格納するドライブを決定する。すなわち、ダミーデータの格納されている複数のドライブ群の中で、どのドライブ内のダミーデータの物理的なアドレスに格納するかを判定する。

【0020】本発明の他の特徴によれば、論理グループ内のあるドライブに障害が発生した場合、正常な残りのドライブ群内のデータとパリティから、障害が発生したドライブ内のデータまたはパリティを回復処理により復元し、この復元したデータまたはパリティを、予備の書き込み領域に書き込む制御を行う。

【0021】

【作用】本発明では、1つのパリティグループを構成する複数のデータやパリティ及び複数のダミーデータを、各々独立したドライブに格納する。

【0022】そのため、書き込みデータ及び書き込み後の新しいパリティを、ダミーデータが格納されていたドライブ群に書き込むことが可能となり、書き込み後のパリティを最小の回転待ち時間で処理できるので、書き込み処理を効率良く処理できる。従来のアレイディスクでは書き込み時に平均1.5回転の回転待ち時間を必要としたのが、平均1回転の回転待ち時間で済むようになる。

【0023】また、常に複数の予備領域が別々のドライブに確保される構成であるので、ドライブの増設時のアドレス制御や、障害発生時の回復処理も簡単になる。

【0024】

【実施例】

【実施例1】以下本発明の一実施例を図1により説明する。本実施例はCPU1、アレイディスクコントローラ(以下ADC)2、アレイディスクユニット(以下ADU)3により構成される。ADU3は複数の論理グループ10により構成され、個々の論理グループ10はm台の磁気記録媒体を用いたSCSI (Small Computer System Interface) ドライブ12と、各々のSCSI ドライブ12とADC2を接続する4つのドライブバス9によ

り構成される。なお、このSCSIドライブ12の数は本発明の効果を得るには、特に制限は無い。

【0025】この論理グループ10は障害回復単位で、同一の論理グループ内のいずれかのドライブが障害が発生したとき、そのグループ内の他のドライブ内のデータを利用して障害が発生したドライブ内のデータを回復するようになっている。

【0026】従来のレベル5のディスクアレイでは、各論理グループは、複数のデータを分散して書き込むためのドライブと、それらのデータから生成したパリティを格納するためのドライブからなる。それぞれらのデータとパリティとの組はパリティグループと呼ばれる。従って、各論理グループは、分散して書き込むデータの数（これをNとする）プラス1のドライブからなる。

【0027】各ドライブは、データ格納用かパリティ格納用ドライブかが定められていた。しかし、本実施例では、各論理グループは少なくともN+3のドライブからなる。したがって、 $m=N+3$ である。2つのドライブは、同一のパリティグループ内のデータのその後の更新時に使用する予備の領域（ダミーデータ領域）を提供するの

【0028】例えば、図3（a）に例示するように、SCSIドライブ#1～#6により1つの論理グループが構成されているとき、データ1～3とパリティP1からなる第1のパリティグループは、ドライブ#1～3、#5に書き込まれると共に、このパリティグループに対して、予備スペースがドライブ#4、6に確保される。本実施例では、これらのダミー領域を用いてデータの更新時のドライブの回転待ち時間を減少させるのが特徴である。すなわち、後に詳述するように、このパリティグループ内のデータ例えば、データ1を更新するとき、このデータ1とパリティP1を読み出し、これらと更新用のデータとから新パリティを生成し、この新パリティと更新用データとをドライブ#4、#6という旧データ1と旧パリティP1を格納していたドライブと異なるドライブ上のダミー領域に書き込む。

【0029】旧データ1と旧パリティP1を保持していた2つの領域は、新たなこのパリティグループのための新たなダミー領域として使用されることになる。以下、実施例の詳細を説明する。

【0030】なお、以下の実施例では、パリティコードを用いた例を説明するが、パリティコードに代えて他のハミングコードないしは誤り訂正コードを用いてもよいことは言うまでもない。

【0031】次にADC2の内部構造について図1を用いて説明する。ADC2はチャンネルバスディレクタ5と2個のクラスタ13とバッテリバックアップ等により不揮発化された半導体メモリであるキャッシュメモリ7により構成される。チャンネルバスディレクタ5は、CPU1から発行されたコマンドを取り込む。このキャッシュ

メモリ7にはドライブ12に書き込まれ又はそこから読み出されたデータと後述するアドレス変換用テーブル40（図3）が格納されている。このキャッシュメモリ7およびその中のアドレス変換用テーブルはADC2内の全てのクラスタにおいて共有で使用される。クラスタ13はADC2内において独立に動作可能なバスの集合で、各クラスタ13間においては電源、回路は全く独立となっている。各クラスタ13にはチャンネルバスディレクタ5とキャッシュメモリ7間のバスとして2つのチャンネルバス6が設けられ、キャッシュメモリ7と、SCSIドライブ12間のバスとして2つのドライブバス6が、それぞれ2個ずつで構成されている。

【0032】それぞれのチャンネルバス6とドライブバス8はキャッシュメモリ7を介して接続されている。CPU1より発行されたコマンドは外部インターフェースバス4を通してADC2のチャンネルバスディレクタ5に発行される。本実施例では、ADC2は2個のクラスタ13により構成され、それぞれのクラスタは2個のバスで構成されるため、ADC2は合計4個のバスにより構成される。このことから、ADC2ではCPU1からのコマンドを同時に4個まで受け付けることが可能である。

【0033】図2は図1のチャンネルバスディレクタ5と1つのクラスタ13の内部構造を示した図である。図2に示すように、CPU1から外部インターフェースバス4を介してADC2に送られてきたコマンドは、チャンネルバスディレクタ5内のインターフェースアダプタ（以下IF Adp）15の1つにより取り込まれる。各チャンネルバス6内のマイクロプロセッサ（以下MP20）は外部インターフェースバス4の中でコマンドを取り込んでいるバスがあるかを調べ、そのような外部インターフェースバス4がある場合は、MP20はチャンネルバススイッチ16を切り換えてそのバスに取り込まれたコマンドの受け付け処理を行なう。もしそのコマンドが受け付けられない場合は受付不可の応答をCPU1へ送る。17は信号線、18はデータバスである。

【0034】24はキャッシュアダプタ回路（C Adp）であり、キャッシュメモリ7に対するデータの読み出し、書き込みをMP20の指示で行う回路で、キャッシュメモリ7の状態の監視、各読み出し、書き込み要求に対し排他制御を行う回路である。C Adp 24により読み出されたデータは、データ制御回路（DCC）22の制御によりチャンネルインターフェース回路（CH IF）21に転送される。CH IF 21ではCPU1におけるチャンネルインターフェースのプロトコルに変換し、チャンネルインターフェースに対応する速度に速度調整する。27は、データの圧縮回路である。28は、Drive IFであり、SCSIの読み出し処理手順に従って、読み出しコマンドをドライブユニットバス9を介して発行する。Drive IF 28では転送されてきた当該データをSCSIドライブ側のキャッシュ

ダプタ回路(C Adp)14に伝送し、(C Adp)14ではキャッシュメモリ7にデータを格納する。35はデータバスである。36は、パリティ生成回路であり、キャッシュメモリ7のデータからパリティを生成し、キャッシュメモリに格納する。127は、回転位置検出回路である。

【0035】本実施例では、レベル5のディスクアレイを実現するので、CPU1からの複数のI/Oコマンドで指定されるデータをまとめて、1つのパリティグループに属するデータを作成し、これらをいずれかのMP20の制御下で、同一の論理グループに書き込む。以下では、説明を簡明にするために、これらのI/Oコマンドで指定されるドライブ番号は同一とする。この際、これらのデータに対するパリティもMP20により生成し、そのパリティグループに属するパリティとして同じ論理グループに属する他のドライブに書き込む。

【0036】図3は、1つの論理グループが6つのドライブからなる場合について、パリティグループに属するデータがどのように書き込まれるかを示すものである。ここでは、同一のパリティグループに属するデータは3つからなり、これらのデータとこれらのデータから生成されたパリティとが1つのパリティグループを形成する。さらに、このグループに対して2つのダミー領域が確保される。これらはいずれも異なるドライブに属する。すなわち、論理グループは、パリティグループとダミーデータ用に使用されるドライブからなる。

【0037】このパリティグループのデータ書き込みが最初に行われるときに、MP20はいくつかの論理グループ内での同一のドライブ内アドレスを有する空領域をサーチし、そこにこれらの3つのデータとパリティと2つのダミーデータ用の領域を別々のドライブに確保する。この際、確保された領域のアドレスに関する情報は後に詳述されるアドレス変換テーブル40に書き込まれる。

【0038】以上の書き込み動作において従来と異なるのは、あるパリティグループを最初に論理ドライブに書き込むときに、そのパリティグループ用の少なくとも2つの予備領域を確保することにある。

【0039】本実施例では、ADU3を構成するSCSIドライブ12は、SCSIインターフェースのドライブを使用する。CPU1をIBMシステム9000シリーズのような大型汎用計算機とした場合、CPU1からはIBMオペレーティングシステム(OS)で動作可能なチャネルインターフェースのコマンド体系にのっとってコマンドが発行される。そこで、SCSIドライブ12をSCSIインターフェースのドライブを使用した場合、CPU1からのコマンドを、SCSIインターフェースのコマンド体系にのっとったコマンドに変換する必要がある。この変換はコマンドのプロトコル変換と、アドレス変換に大きく分けられる。以下にアドレス変換

について説明する。

【0040】CPU1から指定されるアドレスは、図22に示したように、要求データが格納されている当該ドライブ12の番号(ドライブ番号)と、当該ドライブ内のシリンダ番号であるシリンダアドレス(CC)と、シリンダにおいてトラックを選択するヘッド140の番号であるヘッドアドレス(HH)と、レコードアドレス(R)からなる、CCHHRである。

【0041】従来のCKDフォーマット対応の磁気ディスクサブシステム(例えばIBM3990-3390)ではこのアドレスに従ってドライブへアクセスすれば良い。

【0042】しかし、本実施例では、複数のSCSIドライブ12により従来のCKDフォーマット対応の磁気ディスクサブシステムを論理的にエミュレートする。つまり、ADC2は複数のSCSIドライブ12が、従来のCKDフォーマット対応の磁気ディスクサブシステムで使用されているドライブ1台に相当するようにCPU1にみせかける。このため、CPU1から指定してきたアドレス(CCHHR)をSCSIドライブのアドレスにMP20は変換する。

【0043】このアドレス変換には図3の(b)に示すようなアドレス変換テーブル40(以下アドレステーブルとする)が使用される。ADC2内のキャッシュメモリ7には、その内部の適当な領域に、アドレステーブル40が格納されている。本実施例では、CPU1が指定してくるドライブは、CKDフォーマット対応の単体ドライブである。しかし、本発明ではCPU1は単体と認識しているドライブが、実際は複数のSCSIドライブにより構成されるため、論理的なドライブ群として定義される。このため、ADC2のMP20はCPU1より指定してきたアドレス(CPU指定ドライブ番号41とCCHHR46)を1組の論理グループを構成する。SCSIドライブ12の各々に対するSCSIドライブ番号とそのSCSIドライブ内のアドレス(以下SCSI内Addrとする)に変換する。図1に示したように、本実施例では、複数の論理グループを実現できる。アドレステーブル40は各論理グループに対応して生成される。図3(b)には、図3(a)に示したように2つの論理グループがSCASIDライブ#1~#6からなると仮定し、その論理グループに対応するアドレステーブルを示す。

【0044】アドレステーブル40はCPU指定ドライブ番号カラム41とSCSIドライブアドレスカラム42とSCASIDライブ情報カラム43により構成される。これらのカラムの同一行には、同一のパリティグループに属する情報が確保される。SCSIドライブアドレスカラム42はSCSIドライブ内の実際にデータまたはパリティが格納されているアドレスである、SCSI内Addrを格納するカラム44と、パリティが格納

されているSCSIドライブ番号（パリティドライブ番号）を格納するカラム50と、ダミーデータが格納されているSCSIドライブ番号（ダミードライブ番号）を格納するカラム51により構成されている。

【0045】SCASIドライブ情報カラム43は、同一の論理グループに属するSCASIドライブの各々に対応したカラムからなり、それぞれのカラムには、そのドライブに保持されているデータに関連した情報として、CPU1から指定された、ドライブ内アドレスCCHHR6と、そのSCASIドライブ内のデータがキャッシュメモリ7内にも存在する場合の、そのデータのキャッシュメモリ7内のアドレス47と、キャッシュメモリ7がそのデータを保持している場合オン（1）が登録されるキャッシュフラグ48と、そのドライブにダミーデータが格納されている場合オン（1）となる無効フラグ49により構成される。なお、図3ではカラム43内の各カラムには、図3と対比可能なように、それぞれのカラムのそれぞれのエントリがどのデータ及びパリティに関する情報を持つかをそれらのデータまたはパリティで表示してある。

【0046】このテーブル40に保持された情報のうち、パリティドライブカラム50、ダミードライブカラム51に格納されていないドライブはデータ用のドライブであることがわかる。このテーブル40の各行のデータは、対応するパリティグループが最初に論理グループに書き込まれたときに、MP20により書き込まれる。その後、いずれかのパリティグループ内のデータを更新するI/Oコマンドが実行されたとき、このテーブルのそのパリティグループに対する行に属する情報がMP20により更新される。

【0047】このテーブル40のうち、本実施例で特徴的なのはダミードライブ番号が保持されていることである。これは後述する、データの更新時に利用される。このように作成されたテーブル40は、CPU1から論理ドライブ内のデータをアクセスするごとにMP20により利用される。

【0048】例えば図3の例において、CPU1からCPU指定ドライブ番号が”Drive #1”、CCHHRが”ADR8”のデータに対し要求を発行してきた場合、アドレステーブル40でCPU指定ドライブ番号がDrive #1に等しい行に属するSCASIドライブ情報カラム43内エントリを調べ、CCHHRが”ADR8”に等しいエントリを探す。図3の例においては、Data #23に対するCCHHRが”ADR8”となっており、Data #23が要求されたデータであることがわかる。このData #23の物理的地址はアドレステーブル40の同じ行からSCSIドライブ#2内のSCSI内Addr”DADR8”であることがわかる。こうして物理的地址への変換がこのテーブル42よりなされることがわかる。

【0049】このアドレス変換の後、MP20によりSD#2のSCSIドライブ12のデータ#23に対し読み出しまたは書き込み要求が発行される。この時アドレステーブル40においてData #23はキャッシュフラグ48がオフ（0）のため、このデータはキャッシュメモリ7内のCADR2、1に存在しない。もし、キャッシュフラグ48がオン（1）であればキャッシュメモリ7内には、当該データが存在するので、ドライブ12からの上記読み出しまたは書き込みは行わず、キャッシュメモリ内のそのデータに対して行われる。また、このデータは無効フラグ49がオフ（0）のため、このデータはダミーデータでないことがわかる。

【0050】次に、ADC2内での具体的なI/O処理について図1、図2を用いて説明する。CPU1より発行されたコマンドはIF Adp15を介してADC2に取り込まれ、MP120により読み出し要求が書き込み要求が解釈される。

【0051】まず、書き込み時は以下のように処理される。但し、以下ではすでにデータが書き込まれた後に行われるデータ更新のためのデータ書き込みについて述べる。

【0052】CPU1からCPU指定ドライブ番号としてドライブ#1でCCHHRとして”ADR8”を指定するデータ更新コマンドが発行されたとする。まず、ADC2のMP20は、CPU1からこのコマンドを受け取った後、コマンドを受け取ったMP20が所属するクラス13内の各チャンネルバス6において処理可能かどうかを調べ、可能な場合は処理可能だという応答をCPU1へ返す。CPU1では処理可能だという応答を受け取った後にADC2へ書き込みデータを転送する。この転送に先立ち、ADC2ではMP20の指示により、チャンネルバスディレクタ5において、チャンネルバススイッチ16がこのコマンドが転送された外部インターフェースバス4とIF Adp15をチャンネルバス6に接続しCPU1とADC2間の接続を確立する。

【0053】CPU1から転送されてきた書き込みデータ（以下新データまたはNew Dataとする）はMP20の指示により、CHIF21によりプロトコル変換を行ない、外部インターフェースバス4での転送速度からADC2内での処理速度に速度調整する。CHIF21におけるプロトコル変換および速度制御の完了後、データはDCC22によるデータ転送制御を受け、C Adp24に転送され、C Adp24によりキャッシュメモリ7内に格納される。この時、CPU1から送られてきた情報が、CPU指定アドレスの場合は、読みだしと同様にアドレステーブル70によりアドレス変換を行い、物理アドレスに変換する。また、CPU1から送られてきた情報がデータの場合は、キャッシュメモリ7に格納したアドレスを上記アドレス変換により変換した物理アドレスをキャッシュメモリ7に登録する。この

時、対応するキャッシュフラグ48をオン(1)とする。

【0054】この様にキャッシュメモリ7に新データを格納したのをMP20が確認したら、MP20は書き込み処理の完了報告をCPU1に対し報告する。

【0055】なお、キャッシュメモリ7内に保持されている新データに対し、さらに書き込み要求がCPU1から発行された場合は、キャッシュメモリ7内に保持されている新データを書き替える。

【0056】キャッシュメモリ7に新データが格納された後は、以下のようにして新しくパリティを更新し、論理グループ10内のSCSIドライブ12へ新データと新パリティが格納される。

【0057】図2に戻って、MP20はアドレステーブル40を参照したときに、データ、ダミーデータ、パリティが格納されているSCSIドライブ番号を認識する。

【0058】MP20は、Drive1F28に対し、該当するドライブ12に対する書き込み処理を行なうように指示する。

【0059】ここでの書き込み処理は、論理グループ10において新データを実際に書き込む処理と、新データの書き込みによりパリティを新たに作り直す(以下新パリティ、New Parityとする)ための処理すなわち、このパリティを作成するための書き込み前のデータ(以下旧データ、Old Dataとする)と書き込み前のパリティ(以下旧パリティとする)を読み出し、新パリティを作成する処理と、新データと新パリティを書き込む一連の処理からなる。

【0060】図4に示すようにCPU1からSD#1のSCSIドライブ12のデータ#1に書き込み要求が発行された場合、MP20は、SD#1ドライブが所属する論理グループ10において、ドライブSD#1とダミーデータが格納されているドライブSD#4、6とパリティが格納されているドライブSD#5に対し、使用権の獲得を行なう。

【0061】その後は、図5のフローチャートに示すような処理を行なっていく。まず、MP20は回転位置検出回路(RPC)127によりドライブSD#1、4、5、6における回転位置を検出する(ステップ502)。RPC127により回転位置を検出した後にMP20は各SCSIドライブの回転待ち時間を算出する(504)。

【0062】図6に示すように、論理グループ10を構成する各ドライブSD#1~SD#6には旧データ(Data#1~#3)及びダミーデータ(Dummy)、旧パリティ(Parity#1)が格納されている。この論理グループ10内のデータ100は、ディスク130上では図6の下段に示すようになっており、各ドライブSD#1~SD#6によって当該データ100がヘッ

ド140の下に来るまでのディスク130の回転を待つ時間が異なっている。このディスク130の回転を待つ時間のことを回転待ち時間(TW)という。

【0063】回転待ち時間TWを算出した後に、以下のように新パリティと新データを格納するSCSIドライブを決定する。図5のフローチャートに示すように、まず、MP20は旧データ(Old Data)が格納されているSCSIドライブと、旧パリティ(Old Parity)が格納されているSCSIドライブのうち、回転待ち時間TWの大きい方を判定する(506~512)。

【0064】図7の(a)の例では、旧データ(Old Data)が格納されているドライブSD#1の回転待ち時間がTW1、旧パリティ(Old Parity)が格納されているドライブSD#5の回転待ち時間TW5より大きい。そこで、MP20はドライブSD#1のを基準に新データ(New Data)と新パリティ(New Parity)を書き込むドライブを決定する。具体的にはMP20はRPC127からの情報によりダミーデータが格納されているドライブSD#4、6において、ドライブSD#1から書き込み前の旧データ(Old Data)を読み出し(506)、新パリティ(New Parity)の作成後に早く書き込むことが可能な方のドライブを判定し、このドライブに新パリティ(New Parity)を格納することに決定する(512)。一方新データ(New Data)は残りのダミーデータが格納されているドライブに格納することに決定する(514)。このようにして、新データ(New Data)と新パリティ(New Parity)を格納するSCSIドライブ12を決定する。

【0065】図7(a)の例では、ドライブSD#1から書き込み前の旧データを読み出した後は、ドライブSD#6の方がドライブSD#4よりも早く書き込める。(SD#4のSCSIドライブに書き込むためには、1回転待たなければならない。)そのため、ドライブSD#6を書き込み後の新パリティ(New Parity)を格納するドライブとし、ドライブSD#4は、書き込む新データ(New Data)を格納するドライブとする。

【0066】また、図7(b)の例では、ドライブSD#1から書き込み前の旧データ(Old Data)を読み出した後、ドライブSD#4の方がドライブSD#6よりも早く書き込めるため、ドライブSD#4を書き込み後の新パリティ(New Parity)を格納するドライブとし、ドライブSD#6は書き込む新データ(New Data)を格納するドライブとする。

【0067】書き込み後の新データ(New Data)および新パリティ(New Parity)を格納するドライブを決定した後は、図5のフローチャートに従って以下のように処理する。以下の書き込み処理の説

明は、図7の(a)を例に行なう。

【0068】MP20はDrive IF28に新データ(New Data)をドライブSD#4に書き込むよう指示する。Drive IF28ではSCSIの書き込み手順にしたがってドライブユニットパス9の中の1本を介してドライブSD#4に書き込みコマンドを発行する。

【0069】同時にMP20はドライブSD#1に対して旧データ(Old Data)の読み出しと、ドライブSD#5に対して旧パリティ(Old Parity)の読み出しとドライブSD#6への新パリティ(New Parity)の書き込み要求を発行するようにDrive IF28に指示する。

【0070】Drive IF28から読み出したまたは書き込みコマンドを発行されたこれらのドライブでは、Drive IF28から送られてきたSCSI内Addr44へシーク、回転待ちのアクセス処理を行なう。ドライブSD#4においてアクセス処理が完了し書き込みが可能になり次第、C Adp14はキャッシュメモリ7から書き込む新データ(New Data)を読み出してDrive IF28へ転送し、Drive IF28では転送されてきた新データ(New Data)をドライブユニットパス9の中の1本を介してSD#4のドライブへ転送する。新データ(New Data)のドライブSD#4への書き込みが完了すると(510)、ドライブSD#4はDrive IF28に完了報告を行ない、Drive IF28がこの完了報告を受け取ったことを、MP20に報告する(516)。

【0071】SD#4のドライブに対する新データ(New Data)の格納は完了しても、キャッシュメモリ7内には新データ(New Data)が存在しており、パリティの更新はキャッシュメモリ7内に格納されている新データ(New Data)で行なう。

【0072】ドライブSD#1, 5においてもアクセス処理が完了し旧データ(Old Data)および旧パリティ(Old Parity)の読み出しが可能になり次第、旧データ(Old Data)および旧パリティ(Old Parity)を読み出し、キャッシュメモリ7に格納する(512)。ドライブSD#1から旧データ(Old Data)を、ドライブSD#5から旧パリティ(Old Parity)を読み出し、それぞれをキャッシュメモリ7に格納した後、MP20はキャッシュメモリ7内に格納されている書き込む新データ(New Data)とで排他的論理和により、更新後の新パリティ(New Parity)を作成するようにパリティ生成回路(PG)36に指示を出し、PG36において新パリティ(New Parity)を作成しキャッシュメモリ7に格納する(518)。

【0073】新パリティ(New Parity)をキャッシュメモリ7に格納した後、MP120では新パ

リティ(New Parity)の作成が完了したことを認識し、ドライブSD#6に、更新後の新パリティ(New Parity)を書き込むようにDrive IF28に対し指示する。ドライブSD#6への更新後の新パリティ(New Parity)の書き込み方法は、先に述べた書き込む新データ(New Data)をドライブSD#4に書き込んだ方法と同じである。

【0074】ドライブSD#6では既に、Drive IF28から書き込みコマンドが発行されており、指示SCSI内Addr44へシーク、回転待ちのアクセス処理を行なっている(520)。新パリティ(New Parity)は既に作成されキャッシュメモリ7に格納されており、ドライブSD#6におけるアクセス処理が完了した場合、C Adp14はキャッシュメモリ7から新パリティ(New Parity)を読み出してDrive IF28へ転送する。Drive IF28では転送されてきた新パリティ(New Parity)をドライブユニットパス9の中の1本を介してドライブSD#6へ転送する(522)。

【0075】新データ(New Data)及び新パリティ(New Parity)の当該ドライブSD#4, 6への書き込みが完了すると、当該ドライブはDrive IF28に完了報告を行ない、Drive IF28がこの完了報告を受け取ったことを、MP20に報告する。

【0076】この時、MP20は、この新データ(New Data)をキャッシュメモリ7上に残さない場合は、この報告を元にアドレステーブル40のキャッシュフラグ48をオフ(0)にする。さらに、アドレステーブル40において、書き込み前の旧データ(Old Data)の論理アドレス45の無効フラグをオン(1)とし、書き込み前の論理アドレス45内のCCHHR46のアドレスを、書き込み後の論理アドレス45のCCHHR46に登録し、無効フラグをオフ(0)とする。

【0077】また、キャッシュメモリ7内に書き込む新データ(New Data)を保持する場合は、書き込み後のデータに対応するエントリ内のキャッシュアドレス47に、キャッシュメモリ7内の新データ(New Data)が格納されているアドレスに登録し、キャッシュフラグ48をオン(1)とする。さらに、書き込みを行った論理グループ10内のSCSI内Addr44に対し、パリティドライブ番号50とダミードライブ番号51を書き込み後のSCSIドライブ番号43に変更する。

【0078】図7の(b)の例についても同様に処理することが可能である。

【0079】以上述べたようにして書き込み処理を行なうが、その時の回転位置検出方法について以下に詳しく説明する。

【0080】回転待ち時間の算出方法は、図8に示すよ

10

20

30

40

50

うにディスク130の面上をインデックスを基準として扇型に等分割する。本実施例では図8に示すように16分割した。なお、この分割数に制限が無いことは明らかである。この等分割した各領域には番号(1~16)が付けられており、各領域の開始地点にその番号が記録されている。回転待ち時間TWの算出は、各SCSIドライブ12において現在ヘッド140の下にある領域の番号をRPC127に送り、RPC27では当該データの格納されている領域100(アドレスでも可)との差から回転待ち時間TWを計算する。図8を例に計算方法を示す。ヘッド140は領域13の開始地点に位置している。要求データは領域2、3に存在している。要求データの開始領域は領域2である。そこで、ディスク130が一回転する時間を16.6(ms)とすると回転待ち時間TWは以下のように求められる。

【0081】(1)(ヘッドの位置する領域番号-要求データの開始領域番号)>0なら

回転待ち時間=(総分割数-(ヘッドの位置する領域番号-要求データの開始領域番号))/総分割数×一回転時間=(16-(13-2))/16×16.6(ms)

また、要求データの開始領域を領域15とした場合は、

(2)(ヘッドの位置する領域番号-要求データの開始領域番号)≤0、

回転待ち時間=(要求データの開始領域番号-ヘッド位置領域番号)/総分割数×一回転時間=(15-13)/16×16.6(ms)

となる。

【0082】なお、回転待ち時間を時間ではなく、ディスク面上におけるヘッドの位置する領域である回転位置情報そのものから要求データの開始領域までの領域数で考えても問題は無い。

【0083】図9は、図7の(a)の場合での書き込み前の旧データ(Old Data)と旧パリティ(Old Parity)、書き込み後の新データと新パリティのデータの流れを示している。書き込み処理を行なった結果、書き込み前に図10の(A)に示すように、旧データ(Old Data)が格納されていたSD#1のSCSIドライブ12と旧パリティ(Old Parity)が格納されていたSD#5のSCSIドライブ12は、図10の(B)に示すように、書き込み後はダミーデータが格納されているSCSIドライブ12となる。この時、SD#1のSCSIドライブ12には書き込み前の旧データ(Old Data)が格納されており、SD#5のSCSIドライブ12には書き込み前の旧パリティ(Old Parity)が格納されているが、これらは、意味の無いデータまたはパリティでダミーデータとし、書き込み前と書き込み後では論理グループ10においてパリティを作成するパリティグループを構成するSCSIドライブ12が異なる。

【0084】書き込み処理においては、新パリティ(New Parity)を書き込む処理により書き込み処理全体の処理時間が左右される。このように、書き込み後の新データ(New Data)及び新パリティ(New Parity)を、書き込み前の旧データ(Old Data)及び旧パリティ(Old Parity)が格納されているSCSIドライブ12ではなくダミーデータが格納されているSCSIドライブ12に書き込むことにより、データの書き込みにより必要となるパリティの変更を行なった後に、最も早く書き込めるSCSIドライブ12に書き込むことが可能となり、従来のアレイディスクでは図23に示すように書き込み時に平均1.5回転の回転待ち時間が必要としたのが、平均1回転の回転待ち時間で済むため書き込み処理全体の処理時間を短縮できる。

【0085】次に読み出し要求の場合の処理方法を以下に示す。

【0086】MP20が読み出し要求のコマンドを認識すると、MP20はCPU1から送られてきたCPU指定ドライブ番号とCCHHR(以下両方を併せてCPU指定アドレスとする)によりアドレステーブル40を参照し、要求されたデータに対する物理アドレスへの変換を行ない、それとともにそのデータがキャッシュメモリ7内に存在するかどうかをキャッシュフラグ48で判定する。キャッシュフラグ48がオンでキャッシュメモリ7内に格納されている場合(キャッシュヒット)は、MP20がキャッシュメモリ7から当該データを読みだす制御を開始し、キャッシュメモリ7内に無い場合(キャッシュミス)は当該ドライブ12へその内部の当該データを読みだす制御を開始する。

【0087】キャッシュヒット時、MP20はアドレステーブル40によりCPU1から指定してきたCPU指定アドレスをキャッシュメモリ7のアドレスに変換しキャッシュメモリ7へ当該データを読み出しに行く。具体的にはMP20の指示の下で、キャッシュアダプタ回路(CAdp)24によりキャッシュメモリ7から当該データが読み出される。CAdp24はキャッシュメモリ7に対するデータの読み出し、書き込みをMP20の指示で行う回路で、キャッシュメモリ7の状態の監視、各読み出し、書き込み要求に対し排他制御を行う回路である。

【0088】CAdp24により読み出されたデータは、データ制御回路(DCC)22の制御によりチャネルインターフェース回路(CHIF)21に転送される。CHIF21ではCPU1におけるチャネルインターフェースのプロトコルに変換し、チャネルインターフェースに対応する速度に速度調整する。CHIF21におけるプロトコル変換および速度調整の後には、チャネルバスディレクタ5において、チャネルバススイッチ16が外部インターフェースバス4を選択し、IF A

dp15によりCPU1へデータ転送を行なう。

【0089】一方、キャッシュミス時、MP20はアドレス変換テーブル40を用いて、読み出すべきデータが属するドライブとその中のアドレスを判別し、DriveIF28に対し、当該ドライブ12への読み出し要求を発行するように指示する。DriveIF28ではSCSIの読み出し処理手順に従って、読み出しコマンドをドライブユニットバス9を介して発行する。DriveIF28から読み出しコマンドを発行された当該SCSIドライブ12においては、指示されたSCSI内Addr44へシーク、回転待ちのアクセス処理を行なう。当該SCSIドライブ12におけるアクセス処理が完了した後、当該SCSIドライブ12は当該データを読み出しドライブユニットバス9を介してDriveIF28へ転送する。

【0090】DriveIF28では転送されてきた当該データをSCSIドライブ側のキャッシュアダプタ回路(CAdp)14に転送し、(CAdp)14ではキャッシュメモリ7にデータを格納する。この時、CAdp14はキャッシュメモリ7にデータを格納することをMP20に報告する。MP20はこの報告を元にアドレステーブル40内の、CPUが読み出し要求を発行したCPU指定アドレスに対応したエントリ45のキャッシュフラグ48をオン(1)にし、さらにキャッシュアドレス47にキャッシュメモリ7内のデータを格納したアドレスを登録する。キャッシュメモリ7にデータを格納し、アドレステーブル40のこのデータに対応するエントリ45のキャッシュフラグ48をオンにし、キャッシュメモリ7内アドレス47を更新した後は、キャッシュヒット時と同様な手順でCPU1へ当該データを転送する。

【0091】本実施例では、パリティは論理グループ10を構成する各SCSIドライブ12において、各SCSI内Addrが同一のデータに対して作成され、そのパリティもデータと同一のSCSI内Addr44に格納されるとして説明してきた。この方法によれば、ドライブ番号のみでアドレス管理できる利点がある。

【0092】しかし、図10に示すように、パリティ・データが同一SCSI内Addr44ではなく、アドレスが1ずつずれていても同様に実現できる。この場合、各データについて、ドライブ番号のみならずドライブ内アドレスの管理も必要となってくる。システム全体の管理の難易度を考慮して、このずらす量を選定すればよい。

【0093】〔実施例2〕本発明の第2の実施例を図11、図12を用いて説明する。本実施例では、実施例1で示したシステムにおいてSCSIドライブ12に障害が発生した時に、その障害が発生したSCSIドライブ12内のデータを回復し、それを格納するための領域にダミーデータの領域を使用する例を示す。論理グループ

10内の各SCSIドライブ12では、その内部の各々対応する同一SCSI内Addr44のデータによりパリティグループを構成する。具体的には図11に示すように、ドライブSD#1、2、3内のデータ#1、2、3でパリティ生成回路PG36によりパリティ#1が作られ、ドライブSD#5内に格納される。本実施例ではパリティは奇数パリティとし、各々対応するビットについて1の数を数え、奇数であれば0、偶数であれば1とする(排他的論理和)。

【0094】もし、ドライブSD#1に障害が発生したとする。この時、データ#1は読み出せなくなる。

【0095】そこで、図12に示すように、残りのデータ#2、3とパリティ#1をキャッシュメモリ7に転送し、MP20はPG36に対しデータ#1を復元する回復処理を早急に行なうように指示する(1202)。この回復処理を行ないデータ#1を復元した後(1206~1208)、MP120は、アドレステーブル40により、論理グループ内の同一SCSI内Addr44(DADR1)におけるダミーデータの格納されてダミードライブ番号51(SD#4、6)を認識し、このデータ#1をSD#4または6のどちらかのダミーデータのアドレスに格納する(1210~1212)。

【0096】本実施例では、ダミーデータの領域に回復処理により復元されたデータ#1を格納する。これにより、ダミーデータ領域を実施例1で示したような、書き込み処理時の回転待ち時間を短縮させるためだけではなく、SCSIドライブ12に障害が発生したときに、復元したデータを格納するためのスペア領域としても活用する。

【0097】この様に、MP120がダミーデータ領域に回復したデータ#1を格納した後は、キャッシュメモリ7にある図3に示すアドレステーブル40において、ダミードライブ番号51の中で、回復データを格納した方のダミードライブ番号を削除し、この削除したドライブ番号に対するエントリ45に、回復したデータ#1の元のエントリ45の内容を複写する(1214)。

【0098】図11に示すように、ドライブSD#1には、データ#1の他にダミーデータ、パリティ、データ#13、16、19、22が格納されている。ダミーデータについては回復処理を行ない復元する必要は無い。パリティ#3はドライブSD#3、4、5からデータ#7、8、9を読み出して新たに作成しドライブSD#2か6のダミーデータ領域に格納する。データ#13はドライブSD#3、5、6からパリティ、データ#14、15を読み出して、回復処理を行ない復元し、ドライブSD#2または4のダミーデータ領域に格納する。以下同様に他のデータについても回復処理を行ない論理グループ10内のダミーデータ領域に格納していく。

【0099】なお、ドライブSD#2、3、4、5、6のダミーデータ領域に、ドライブSD#1の回復データ

10

20

30

40

50

を全て格納した後は、ダミーデータが論理グループ10において一つしかないため、実施例1で述べたような書き込み時の回転待ちを短くすることは出来ないため、従来のアレイディスクであるRAIDのレベル5の処理となる。また、ドライブSD#1の回復データを全て格納した後は、ドライブSD#2, 3, 4, 5, 6の中で更にもう一台のドライブに障害が発生した場合、同様にその障害が発生したドライブ内のデータについて回復処理を行ない、残りのダミーデータ領域に格納し、処理を行なえる。

【0100】この様にして、論理グループ10内のダミーデータ領域を全て使いきってしまった場合は、障害SCSIドライブ12を正常のSCSIドライブ12に交換し、この交換した正常なSCSIドライブ12は全てダミーデータ領域として論理グループを再構成する。障害SCSIドライブ12を正常のSCSIドライブ12に交換した直後は、ダミーデータ領域が特定SCSIドライブ12に集中した形になっているため、このSCSIドライブ12が使用出来ずにまたされることが多くなりネックとなるため、実施例1で示した回転待ち時間を短縮する効果が、効率的に発揮出来ない。

【0101】しかし、時間が立つにつれて、ダミーデータが分散されて、SCSIドライブ障害前の状態に戻っていき、次第に解消されていく。もし、この時間が問題となる場合は、SCSIドライブに障害が発生したことを感知した場合、正常なSCSIドライブに交換して、この交換した正常なSCSIドライブに障害が発生したSCSIドライブ内のデータとパリティをユーザが復元することも可能とする。なお、この時ダミーデータに関しては復元せずにダミーデータ領域として空けておく。

【0102】本実施例ではこの回復処理と、ダミーデータ領域へ復元したデータを書き込む処理をMP20が自動的に行なう。この様に自動的に行なうことによりSCSIドライブに障害が発生した場合、障害が発生したSCSIドライブを正常なSCSIドライブに交換し回復したデータを書き込むのと比較し、本発明ではシステムを使用するユーザがSCSIドライブに障害が発生するとすぐに正常なSCSIドライブと交換する必要が無いため、ユーザの負担が軽くなる。

【0103】〔実施例3〕本発明の第3の実施例を図13～図15により説明する。本実施例では、実施例1、2において示したディスクアレイシステムの論理グループ10において、要求性能に合わせてダミーデータの数を増やすことを可能とする。ここでいう要求性能とは読み出し／書き込み処理を高速に行ない処理時間を短縮させるアクセス性能と、障害発生時に対処する信頼性能の両方をいう。ダミーデータの数を増加させる方法は、図13のように全てがダミーデータにより構成されている拡張ドライブを増設する。

【0104】一例として図14に示すように、論理グル

ープ10内においてダミーデータの数を3個にした場合を示す。この例では、SD#1, 2, 3のSCSIドライブ12にデータが格納されており、SD#4, 6及びSD#7（拡張ドライブ）のSCSIドライブ12にはダミーデータが格納されており、SD#5のSCSIドライブ12にはパリティが格納されている。

【0105】本実施例でのデータの更新時の処理は、実施例1に比べて、ダミードライブの数が多点異なる。つまり新パリティを書き込むドライブを選択するとき、3つのダミードライブから実施例1の方法で選択する。新データの書き込みは残りの2つのダミードライブのうち、より早く書き込めるドライブを選択する。図14はそのときの動作の一例である。

【0106】本実施例のようにダミーデータを増加させることにより、書き込み処理における選択対象のドライブが増え、書き込み時のオーバーヘッド（回転待ち）を減少させることが可能となる。また、障害が発生した場合、実施例2のようにその障害が発生したデータを回復して、この回復したデータを格納することが可能であり、この回復したデータ格納する領域を増加させることにより信頼性の向上を図ることが可能となる。高性能な処理を要求するデータはこの様にダミーデータを多く持つようにする。また、一つの論理グループ10内においてダミーデータの数が違うグループ混在させることも可能である。

【0107】MP20は当該論理グループ10内において、どのデータ、ダミーデータ、パリティがどの様に構成されているか、キャッシュメモリ7内のアドレステーブル40により認識することが可能である。このため、CPU1から読み出したまたは書き込み要求が発行されてきた場合、当該データが格納されている論理グループ10の構成を、アドレステーブル40を調べることで認識し、実施例1のように処理するか、実施例3のように処理するかを、MP20が判断し、処理を行なう。

【0108】なお、本実施例では実施例2の効果を実現することが可能であることは明かである。また、本実施例ではダミーデータ数を3としたが、この数は本発明を制約しない。

【0109】本実施例では、図13に示すように、拡張ドライブにはダミーデータのみが格納されているとして、論理グループ10における同一SCSI内Addr 44でダミーデータ数を増加した。

【0110】このように、ダミーデータ数を増加させた後、容量または、性能の向上を図るため、ダミーデータの一部に、新規にデータを格納する方法を図15で説明する。図15は、ダミーデータの追加処理（A）と、容量増設に伴う新規データの書き込み処理（B）とからなっている。ダミーデータの追加処理（A）では、図13に示したように、6台のSCSIドライブ12により構成されていた論理グループ内に、1台のSCSIドライ

10

20

30

40

50

ブ12を増設し、これを拡張ドライブとした場合、先に述べたように、この拡張ドライブは、ダミーデータにより構成されているとし、論理グループ10における同一SCSI内Addr44でダミーデータ数を増加させる。この時、アドレステーブル40のSCSIドライブ番号43に拡張ドライブのSCSIドライブ番号を追加し、ダミードライブ番号51に拡張ドライブのSCSIドライブ番号を追加する。また、追加したSCSIドライブ番号の各SCSI内Addr44に対応する全論理アドレス45において無効フラグ49をオン(1)として全てをダミーデータとする(1502)。

【0111】次に新規データの書き込み処理(B)を説明する。もし、ユーザから容量の増設要求がADC2のMP120に発行され、新規データを書き込む場合、MP120はアドレステーブル40により論理グループ10における同一SCSI内Addr44の3個のダミーデータのアドレスを認識し、現在、他の読み出し/書き込み処理を行っていないSCSIドライブ12に格納されてダミーデータを1個選択する(1504)。MP120によるダミーデータの選択が完了したら、MP120は、新規データをこの選択したダミーデータのアドレスに通常の書き込み処理と同様に書き込む(1506)。なお、この時、パリティの更新は、旧データはダミーデータが全て0として、旧データの読み出しは不要である。また、新パリティの書き込みについては、新規データを書き込んだダミーデータ以外の残りの2個のダミーデータの中で、どちらか早く書き込める方に書き込んでみてもかまわない。

【0112】以上のように、ダミーデータに新規データを書き込んだ後、MP120はアドレステーブル40において、新規データが書き込まれたダミーデータの論理アドレス45を新規データに変更し、ダミードライブ番号カラム51から新規データが書き込まれたダミーデータのSCSIドライブ番号を削除する(1508)。以下同様に新規データを次々に書き込み、容量の増設を図る。

【0113】このように、論理グループ10に分散されているダミーデータに新規データを分散して書き込むため、拡張したドライブに対する新規データの書き込みが、集中しないため、効率良く処理することが可能となる。

【0114】〔実施例4〕本発明の第4の実施例を図16で説明する。本実施例では複数の論理グループ10間に渡ってダミーデータを共有することを可能にする。ディスクアレイサブシステムは、図16に示すように複数の論理グループ10により構成される。本実施例では実施例3で述べたような、高性能を要求するデータについてダミーデータを増加させるのを、論理グループ10内のSCSIドライブ12の数を増加させるのではなく、論理グループ10間のSCSIドライブ12の割り当て

方を可変とすることで実現する。本実施例では、サブ論理グループはパリティグループと、このパリティグループに対応するダミーデータにより構成されるグループと定義し、論理グループ10はこのサブ論理グループにより構成されるとする。

【0115】図16においてサブ論理グループ#1とサブ論理グループ#2においてSD#7のSCSIドライブ12のダミーデータを共有する場合を以下に示す。サブ論理グループ#1においてはドライブSD#1のデータ#1、ドライブSD#2のデータ#2、SD#3のデータ#3とSD#4、6のダミーデータとSD#5のパリティにより構成されている。サブ論理グループ#1についてはドライブSD#7のダミーデータをサブ論理グループ#2と共有し拡大論理グループ#1とする。

【0116】拡大論理グループ#1ではダミーデータが3個あり、サブ論理グループ#2ではダミーデータは2個である。一方、サブ論理グループ#2においてはドライブSD#8のデータ#4、SD#10のデータ#5、SD#11のデータ#6、SD#7、9のダミーデータと、SD#12のパリティ#2により構成されている。

【0117】拡大論理グループ#1では実施例2、3での制御が行なわれ、サブ論理グループ#2では実施例1、2の制御が行なわれる。MP20は当該サブ論理グループが、どのSCSIドライブ12内のどのデータとパリティとダミーデータにより構成されているか、キャッシュメモリ7内のアドレステーブル40により認識することが可能である。アドレステーブル40にはCPU1が指定してきたCPU指定アドレスが、どのサブ論理グループに属するものなのかのテーブルを持っている。このため、CPU1から読み出したまたは書き込み要求が発行されてきた場合、アドレステーブル40により当該データが、どのサブ論理グループに所属するか、MP20が認識することが可能である。

【0118】以上のように本実施例ではサブ論理グループ間においてダミーデータを共有し拡大論理グループを構成することが可能となる。拡大論理グループにおいて、どのサブ論理グループがダミーデータを共有するかはユーザは自由に指定可能である。この時、サブ論理グループ間において、CPU1からの読み出しまたは書き込み要求が競合しないような論理グループ10でダミーデータを共有するようにする方が良い。

【0119】なお、各パリティグループごとに予備領域を持つことは、管理を単純化する効果があるが、用途によっては、パリティグループごとに対応させる必要はない。

【0120】すなわち、各パリティグループ対応の予備領域でなくて、複数のパリティグループに使用可能な予備領域を各ドライブに設けておき、新データ、新パリティを書き込むときに修正前のパリティグループが属するドライブ以外の2つのドライブにある2つの予備領域に

新データと新パリティを書き込んでもよい。旧データ、旧パリティは新たに予備領域とする。この方法によれば、予備領域の絶対量が少なくて済むという効果がある。

【0121】〔実施例5〕本発明の第5の実施例を図17～図20により説明する。本実施例では図17、18に示すように、論理グループ10単位にサブDKC11を設ける。そして、その内部に図19に示すように、実施例1、2、3、4において示したキャッシュメモリ7内のアドレステーブル40と、PG36、サブキャッシュ32とそれらを制御するマイクロプロセッサMP329を持たせたものである。本実施例におけるデータの処理手順は実施例1及び2で示したものと同様である。以下には実施例1、2、3、4で示した処理手順と異なる部分のみを説明する。

【0122】本実施例では図19に示すように実施例1、2、3、4で示したキャッシュメモリ7内のアドレステーブル40を、サブDKC11内のデータアドレステーブル(DAT)30に格納する。DAT30は格納されているテーブルの形式や機能は実施例1、2、3、4のテーブル40と同様であるが、異なるのはデータを格納するSCSIドライブアドレスが論理グループ10に属するものに限られている点と、書き込み、読み出しデータは保持しない専用メモリ上に構成されていることである。このため、以下では図3(b)のテーブル40をDAT30の代りに適宜引用する。ADC2内のGAT23はCPU1から指示されたCPU指定アドレスから、そのCPU指定アドレスが指示する場所がADU3内のどの論理グループ10かを判定するのみである。

【0123】キャッシュメモリ7内には、その特別な領域に図20に示すような論理グループテーブル(LGT)60が格納されている。LGT60は図20に示すようにCPU1から指定されるCPU指定ドライブ番号41とCCHHR46に対応して、論理グループアドレス61が決定できるテーブルとなっている。また、LGT60にはCPU指定アドレスに対応するデータがキャッシュメモリ7内に存在する場合、そのデータのキャッシュメモリ7内のアドレスをキャッシュアドレス47に登録でき、また、キャッシュメモリ7内にデータが存在する場合オン(1)とし、キャッシュメモリ7内に存在しない場合オフ(0)とするキャッシュフラグ48が用意されている。ユーザは初期設定する際に自分の使用可能な容量に対する領域を確保するが、その際にADC2のMP120が、LGT60により論理グループ10を割当てて。

【0124】この時、MP20はLGT60にユーザが確保するために指定したCPU指定アドレスに対応する領域を登録する。そこで、実際の読みだし、書き込み処理においては、GAT23はLGT60によりCPU1から指定してきたCPU指定アドレスに対応した論理グ

ループ10を認識することが可能となる。

【0125】読み出し時はGAT23がLGCにより論理グループ10を確定し、その確定した結果をMP20に報告し、MP20はこの当該論理グループ10に対し読み出し要求を発行するようにDriveIF28に指示する。MP20から指示を受けたDriveIF28は当該論理グループ10のサブDKC11に対し読み出し要求とCPU1が指定するCPU指定アドレスを発行する。サブDKC11ではマイクロプロセッサ(MP)29がこの読み出し要求のコマンドとCPU指定アドレスを受け付け、実施例1と同様DriveIF28から送られてきたCPU指定アドレスをDAT30を参照し、当該データが格納されている論理グループ10内の物理アドレスに変換し、この物理アドレスから当該SCSIドライブアドレス(SCSIドライブ番号と其中的SCSI内Addr)を確定する。このアドレスの確定後MP29は当該SCSIドライブ12に対し、読み出し要求を発行する。

【0126】MP29から読み出し要求を発行されたSCSIドライブ12ではそのSCSI内Addrヘシーク、回転待ちを行ない、当該データの読み出しが可能になり次第、該当データをドライブアダプタ回路(DriveAdp)34に転送し、DriveAdp34はサブキャッシュメモリ32に格納する。サブキャッシュメモリ32に当該データの格納が完了した後、DriveAdp34はMP329に格納報告を行ない、MP29はDAT30の当該データの当該論理アドレス45内の当該キャッシュフラグ48をオン(1)とする。後に当該キャッシュフラグ48がオン(1)のデータに対し読み出しまたは書き込み要求が発行された場合は、サブキャッシュ32内で処理を行なう。

【0127】実施例1と同様にMP29によるDAT30の更新が終了すると、MP29はADC2内のDriveIF28に対しデータ転送可能という応答を行ない、DriveIF28はこの応答を受け取ると、MP20に対し報告する。MP20はこの報告を受け取ると、キャッシュメモリ7への格納が可能なら、DriveIF28に対しサブDKC11からデータを転送するように指示する。DriveIF28ではMP20からの指示を受けるとサブDKC11のMP29に対し読み出し要求を発行する。この読み出し要求を受けたMP29はサブキャッシュアダプタ回路(SCA)31に対しサブキャッシュ32から当該データを読みだすように指示し、SCA31は実際にデータを読み出してDriveIF28にデータを転送する。DriveIF28がデータを受け取った後は、実施例1、2で示した処理を行なう。

【0128】一方、書き込み時は読み出し時と同様に当該論理グループ10を確定し、MP20はDriveIF28に対し当該論理グループ10のMP29に対し書

込み要求を発行するように指示する。当該論理グループ10内のMP329は、書き込み要求を受け付け、書き込みデータをサブキャッシュ32に格納した後は、図5のフローチャートに従って、実施例1、2、3、4と同様にサブDKC11内で当該SCSIドライブ12群において回転位置の検出を行なう。MP329は回転位置の検出を行なった後は書き込みSCSIドライブ12を確定する。MP29により書き込みSCSIドライブ12を確定した後は実施例1、2、3、4と同様に処理する。本実施例では実施例1、2、3、4と同様の効果を実現することが可能である。

【0129】以上述べた各実施例においては、磁気ディスク装置を用いたシステムを例に説明したが、本発明は光ディスク装置を用いた記憶システムのような他の回転型の記録媒体に用いても同様な効果を発揮することが可能である。

【0130】

【発明の効果】本発明によれば、データの更新時における回転待ち時間を短縮できるため、書き込み処理を高速に行え、これにより単位時間当りのI/O処理件数を増加させることが可能となる。また、SCSIドライブに障害が発生した場合、すぐに正常なSCSIドライブと交換する必要が無いため、システムを使用するユーザの負担が軽くなる。また、アドレス管理が簡単なため、ドライブの拡張も容易である。さらに、通常は使用しないスペアのSCSIドライブを、回転待ち時間の短縮という、性能向上のために使用でき、SCSIドライブ資源の有効活用が図れる。

【図面の簡単な説明】

【図1】本発明の第1の実施例の全体構成図。

【図2】第1の実施例のクラスタ内構成図。

【図3】アドレス変換テーブルの説明図。

【図4】更新処理時のデータ移動説明図。

【図5】更新処理フローチャート。

【図6】パリティグループを構成する各データのディスク上での位置説明図。

【図7】更新処理タイミングチャート。

【図8】ディスクの回転位置検出方法説明図。

【図9】更新前後におけるパリティグループ内データの詳細説明図。

【図10】第1の実施例の変形例になる論理グループのアドレス構成図。

【図11】第2の実施例におけるデータ回復処理説明図。

【図12】図11の例の障害回復処理のフローチャート。

【図13】第3の実施例におけるドライブ拡張例の構成図。

【図14】図13の例における更新処理タイミングチャート。

【図15】ドライブ拡張時のデータ書き込み処理フローチャート。

【図16】第4の実施例の論理グループの説明図。

【図17】第5の実施例のシステム構成図。

【図18】第5の実施例のクラスタ内構成図。

【図19】第5の実施例のサブDKC内構成図。

【図20】第5の実施例の論理グループテーブル説明図。

【図21】従来の汎用大型計算機とRAID Level 15における更新処理説明図。

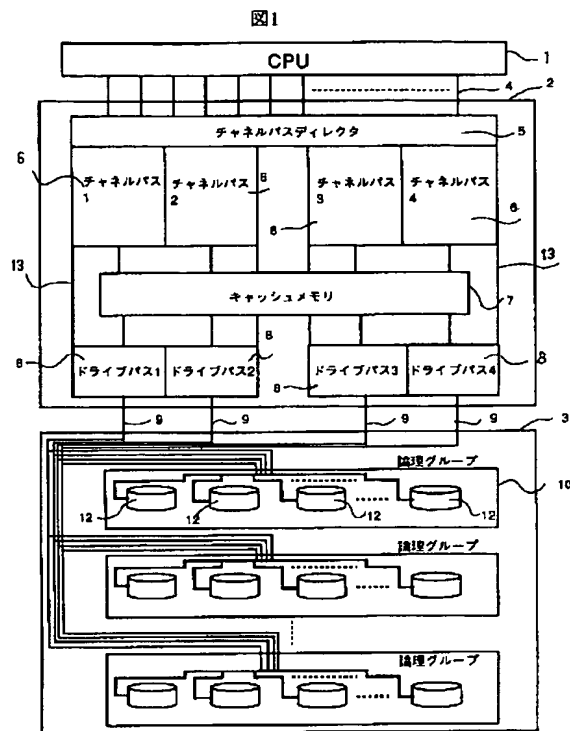
【図22】図21の例におけるドライブ内アドレス説明図。

【図23】図21のRAID 5における書き込み処理タイミングチャート。

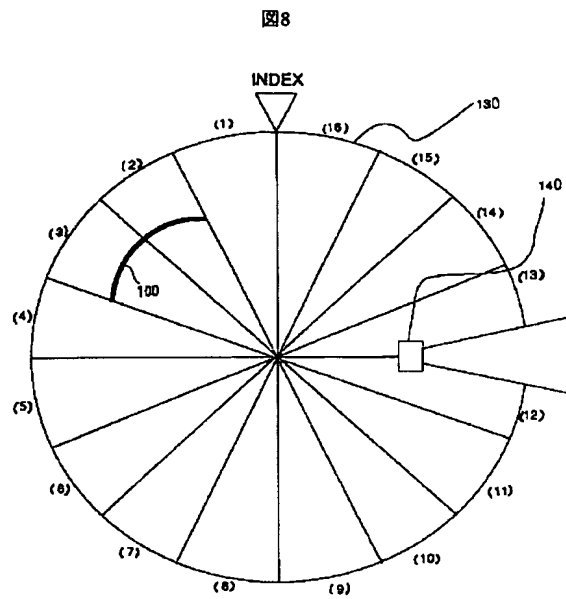
【符号の説明】

1…CPU、2…アレイディスクコントローラ(ADC)、3…アレイディスクユニット(ADU)、4…外部インターフェースバス、5…チャンネルバスディレクタ、6…チャンネルバス、7…キャッシュメモリ、8…ドライブバス、9…アレイディスクユニットバス、10…論理グループ、12…SCSIドライブ、13…クラスタ、14…ドライブ側キャッシュアダプタ(CAdp)、15…インターフェースアダプタ、16…チャンネルバススイッチ、17…制御信号線、18…データ線、19…バス、20…マイクロプロセッサ1(MP1)、21…チャンネルインターフェース(CHIF)回路、22…データ制御回路(DCC)、23…グループアドレス変換回路(GAT)、24…チャンネル側キャッシュアダプタ(CAdp)、28…ドライブインターフェース回路(DriveIF)、29…マイクロプロセッサ3(MP3)、30…データアドレステーブル40、31…サブキャッシュアダプタ、34…ドライブアダプタ(DriveAdp)、35…ドライブバス、36…パリティ生成回路、40…アドレス変換用テーブル(アドレステーブル)、47…キャッシュアドレス、48…キャッシュフラグ、49…無効フラグ、60…論理グループテーブル、61…論理グループアドレス

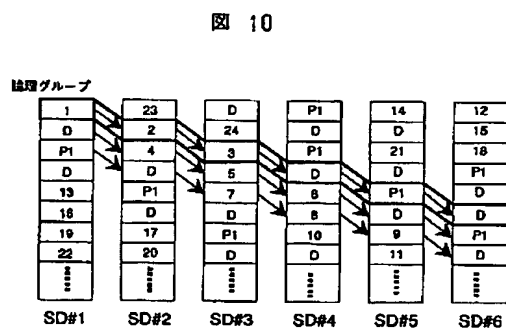
【図1】



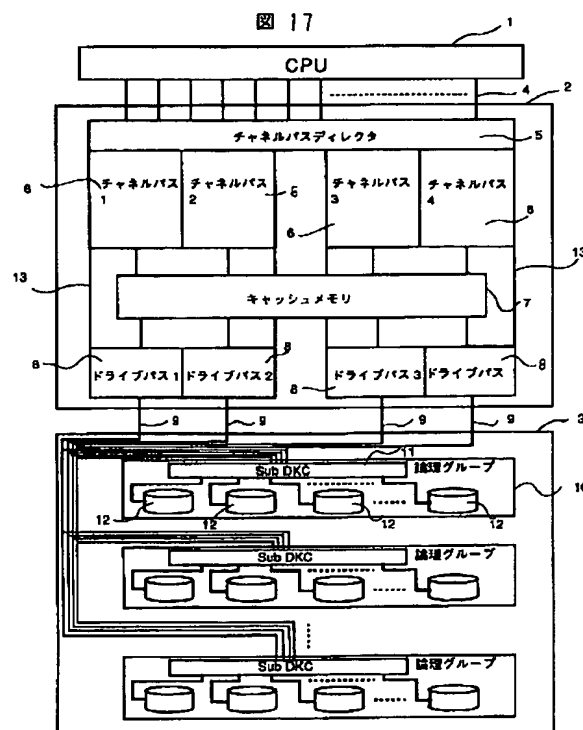
【図8】



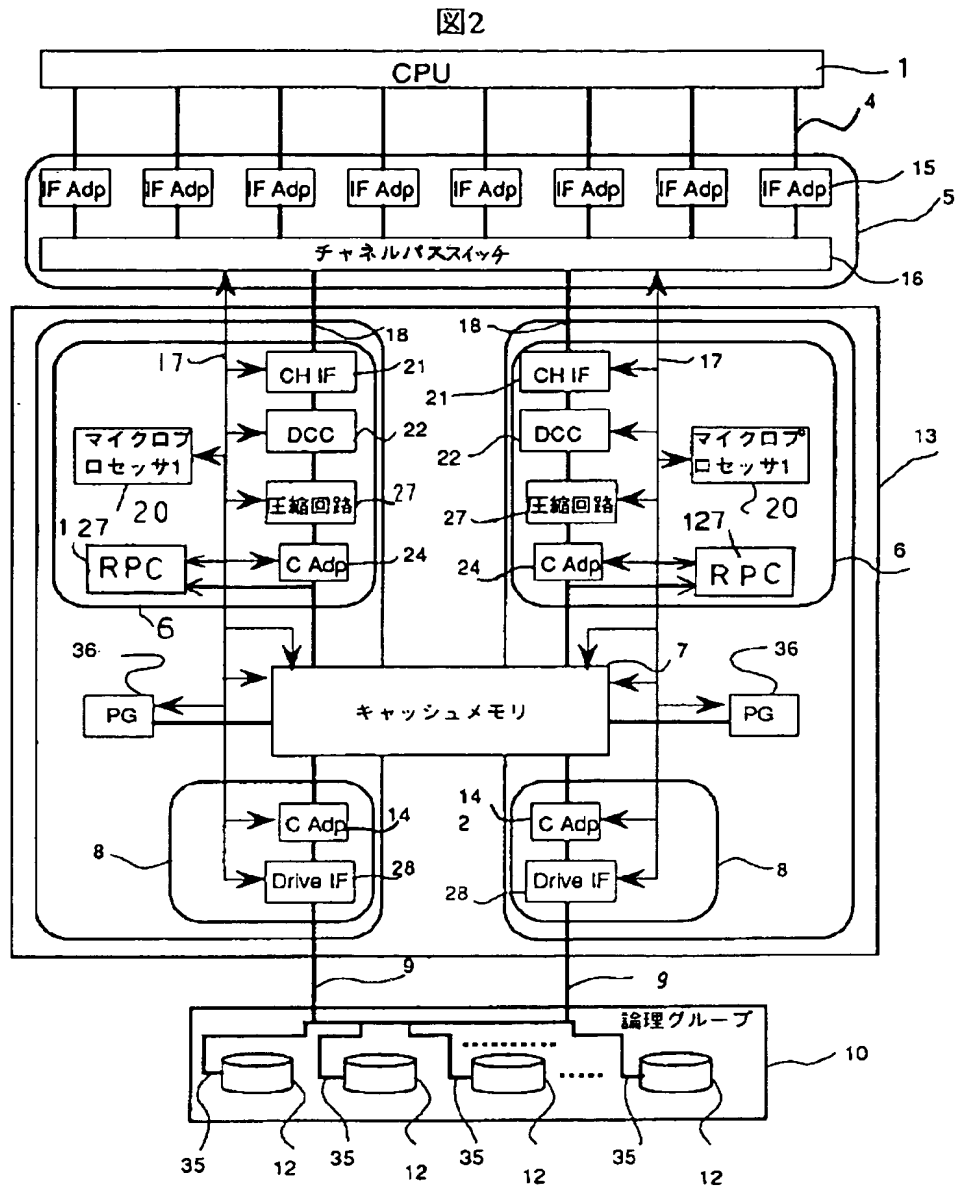
【図10】



【図17】

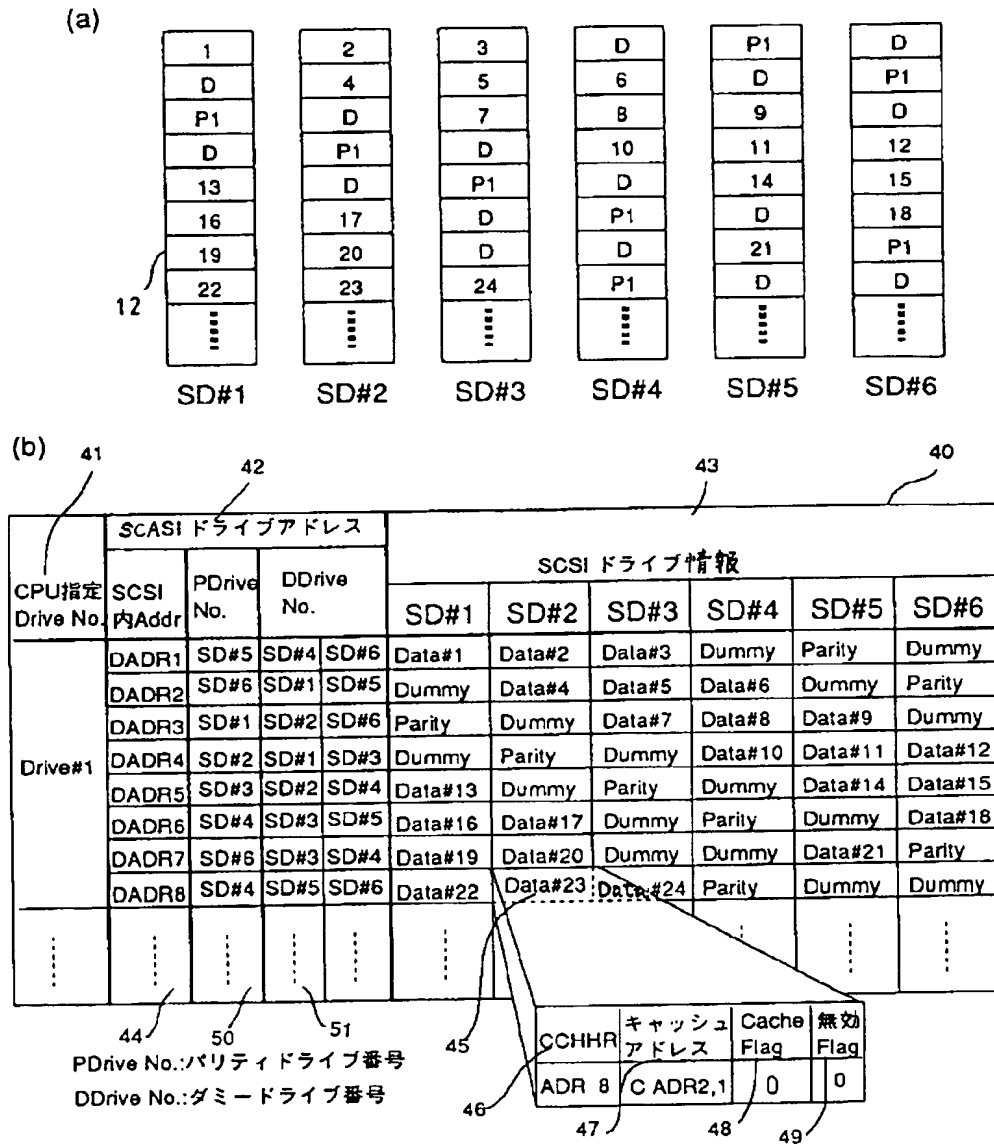


【図2】



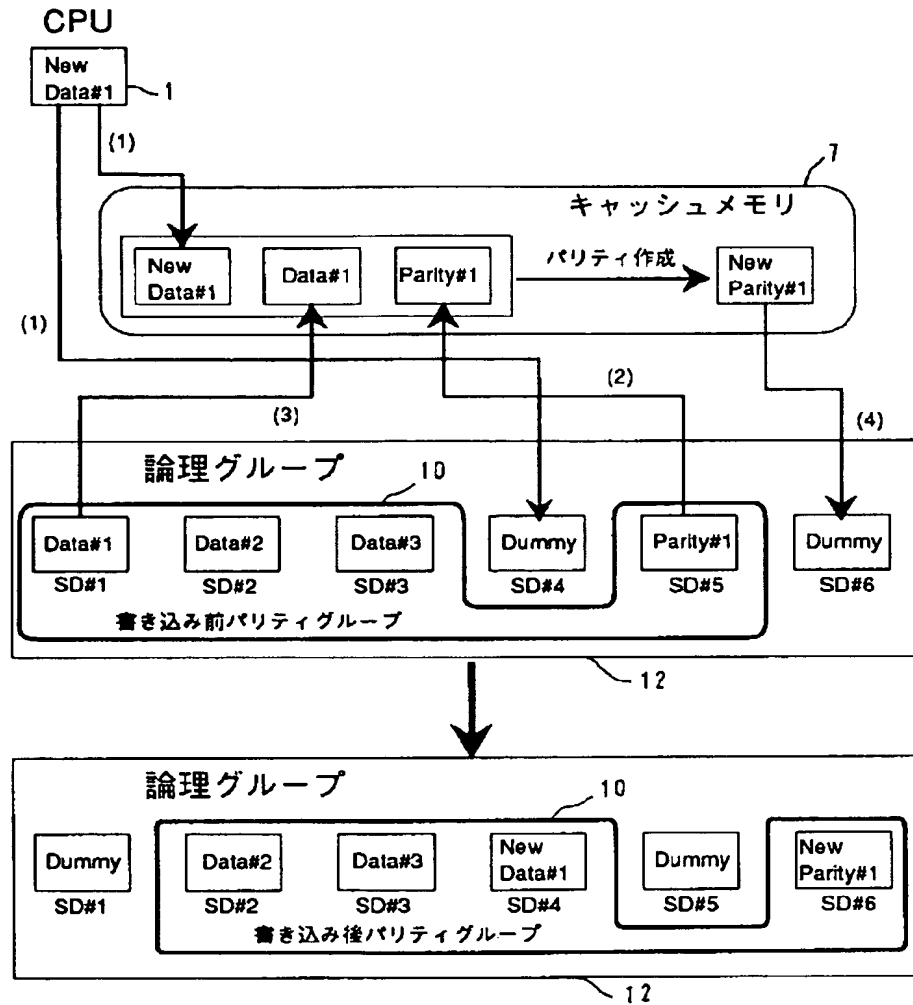
【図3】

図3



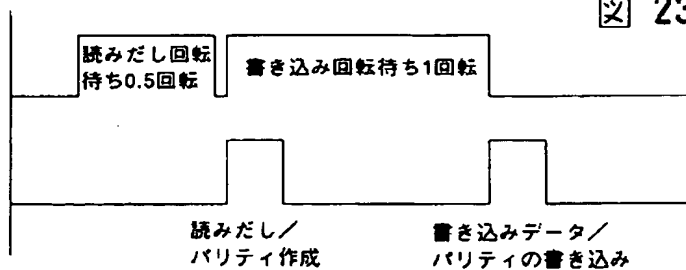
【図4】

図4

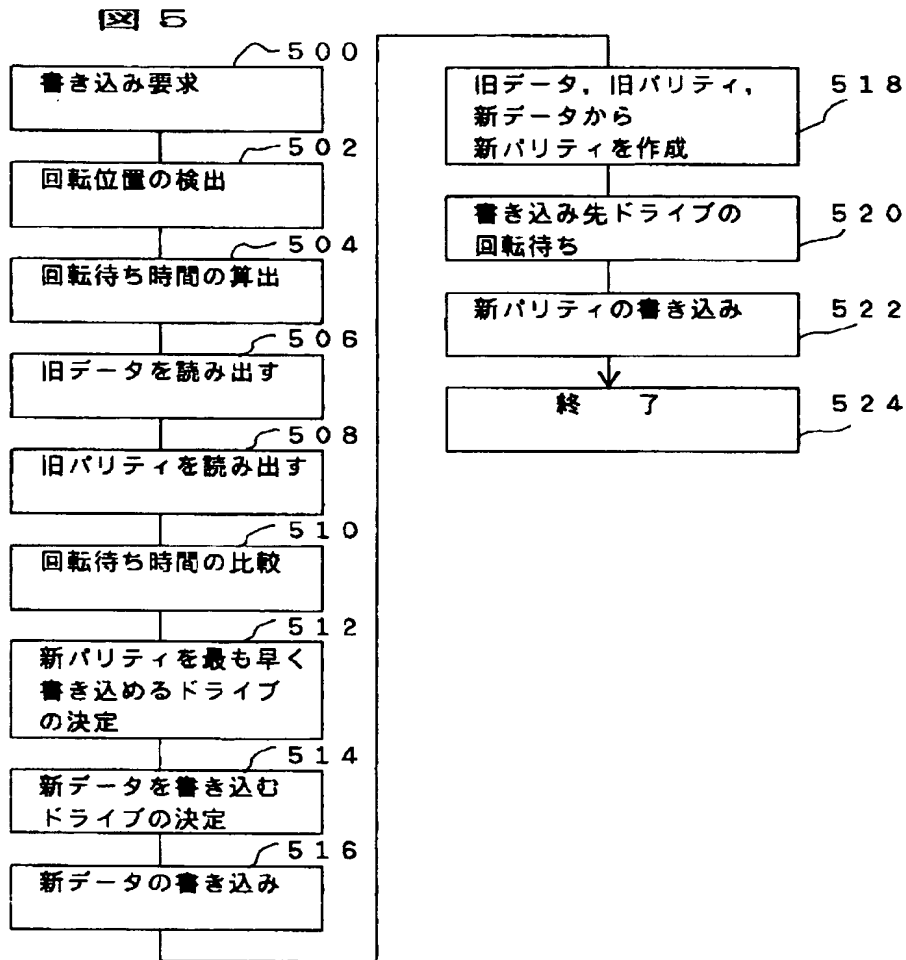


【図23】

図 23

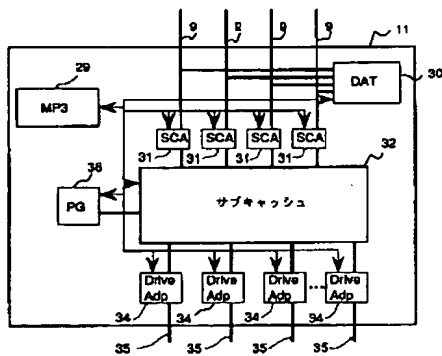


【図5】



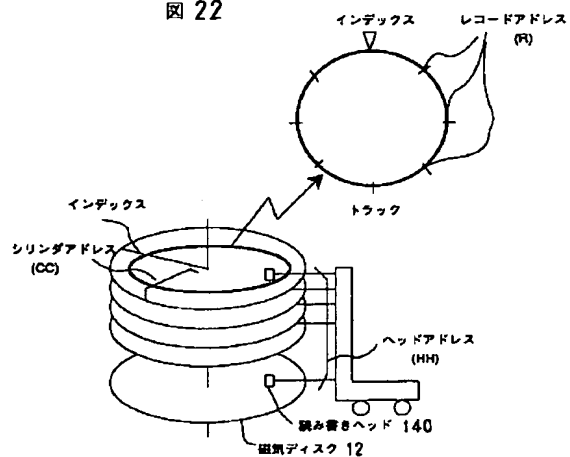
【図19】

図19



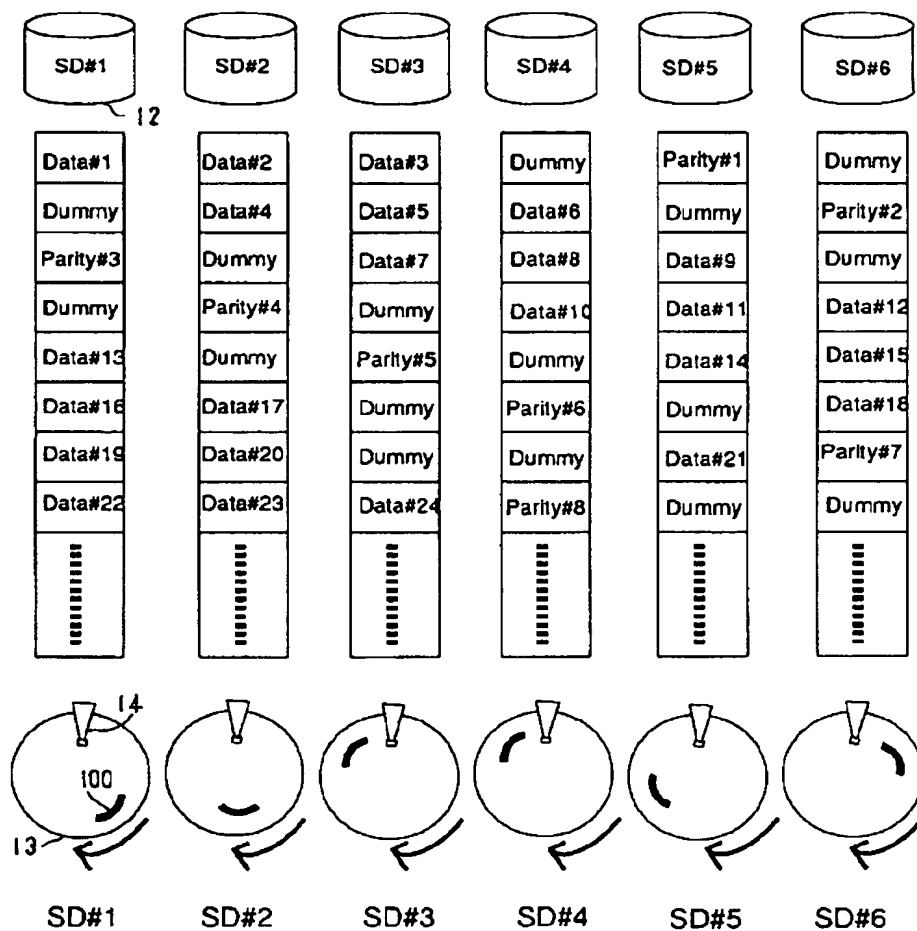
【図22】

図22



【図6】

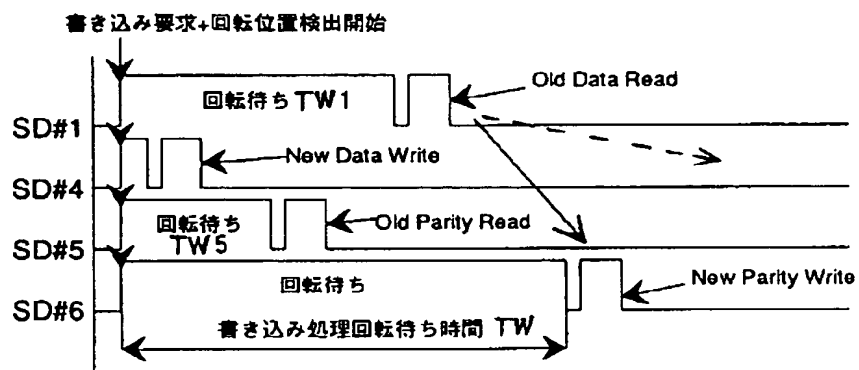
図6



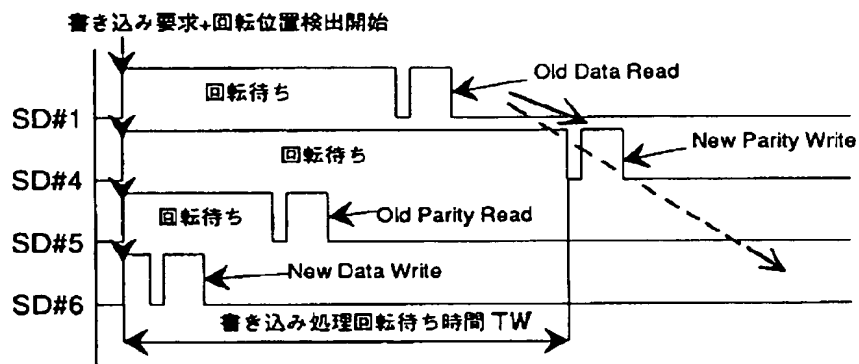
【図7】

図7

(a)



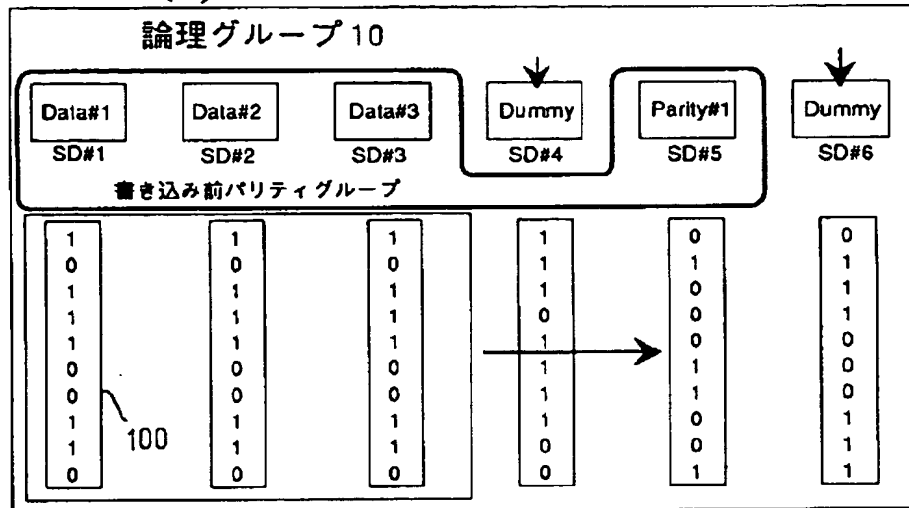
(b)



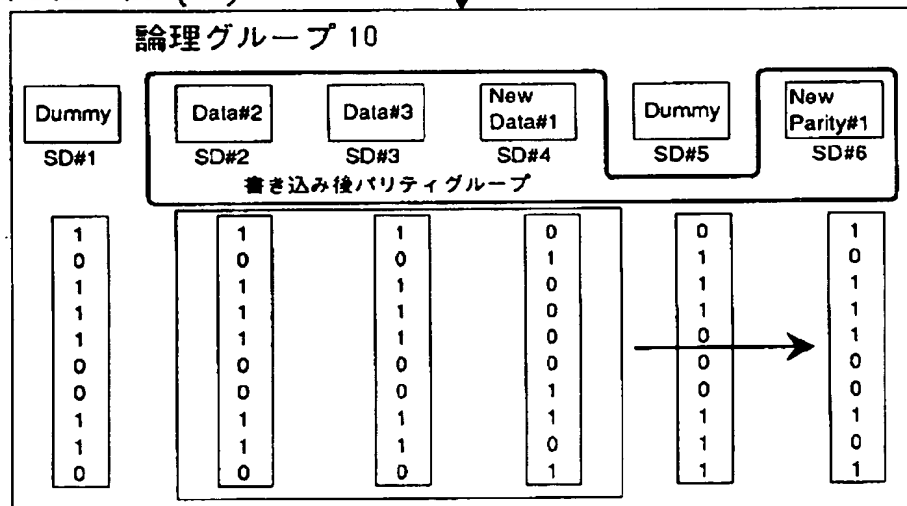
【図9】

図9

書き込み前 (A)

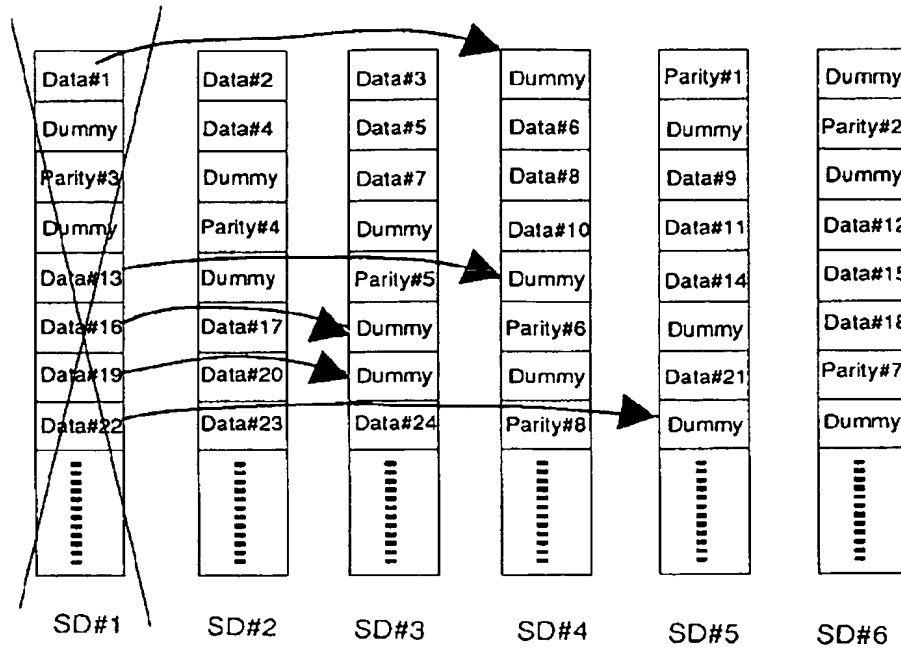


書き込み後 (B)

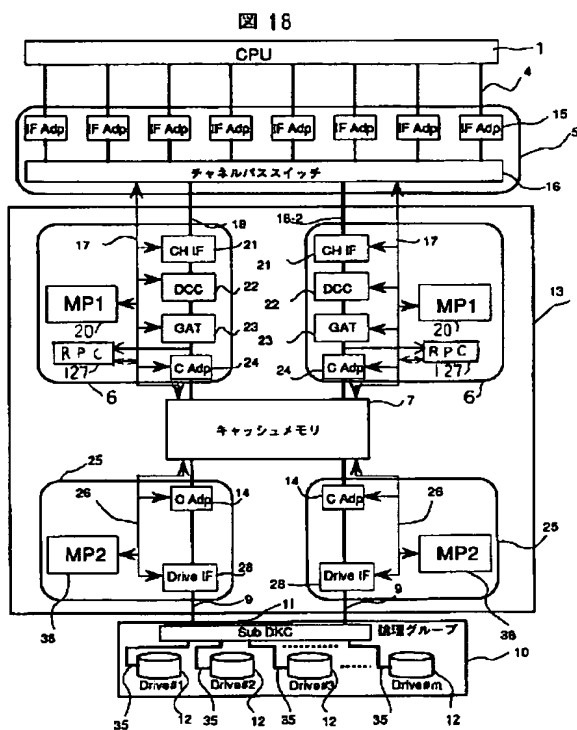


【図11】

図 11



【図18】



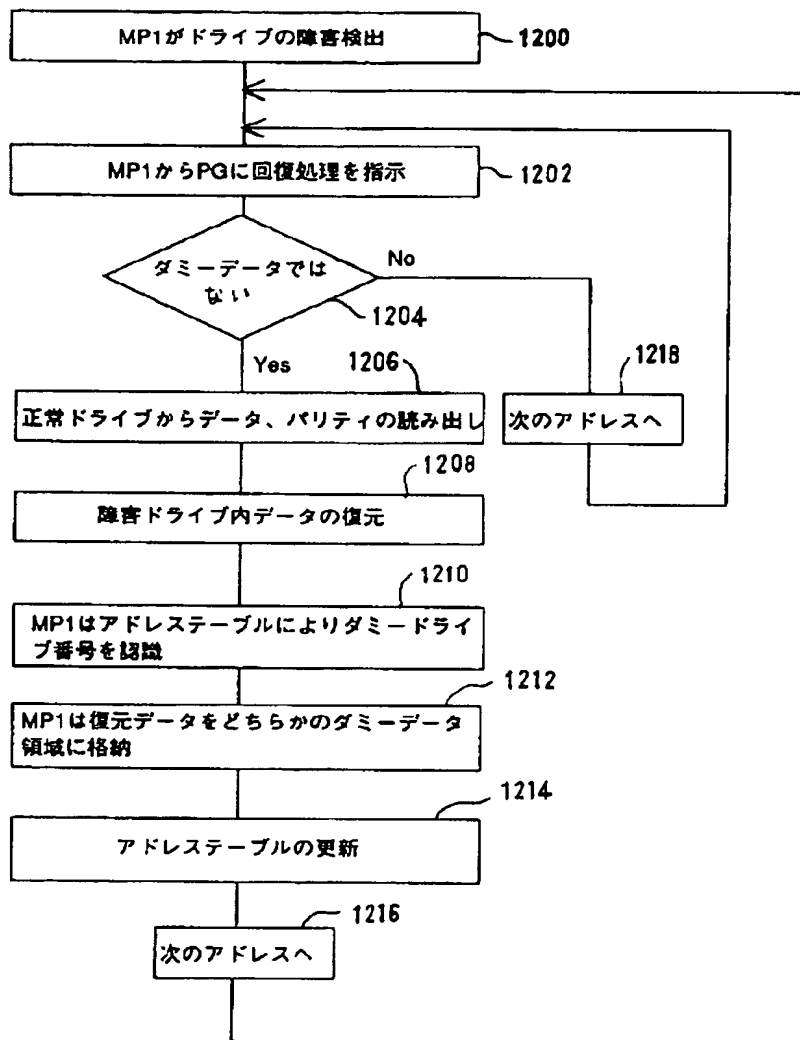
【図20】

図 20

CPU指定アドレス				
CPU指定 Drive No.	CCHHR	論理グループ アドレス	キャッシュ アドレス	Cache Flag
Drive#1	ADR 1	LADR 1	—	—
	ADR 2	LADR 3	—	—
	ADR 3	LADR 8	—	—
	⋮	⋮	⋮	⋮
Drive#2	ADR 1	LADR 2	—	0
	ADR 2	LADR 1	CADR1,5	1
	ADR 3	LADR 5	CADR1,8	1
	ADR 4	LADR 4	CADR1,6	1
⋮	⋮	⋮	⋮	⋮

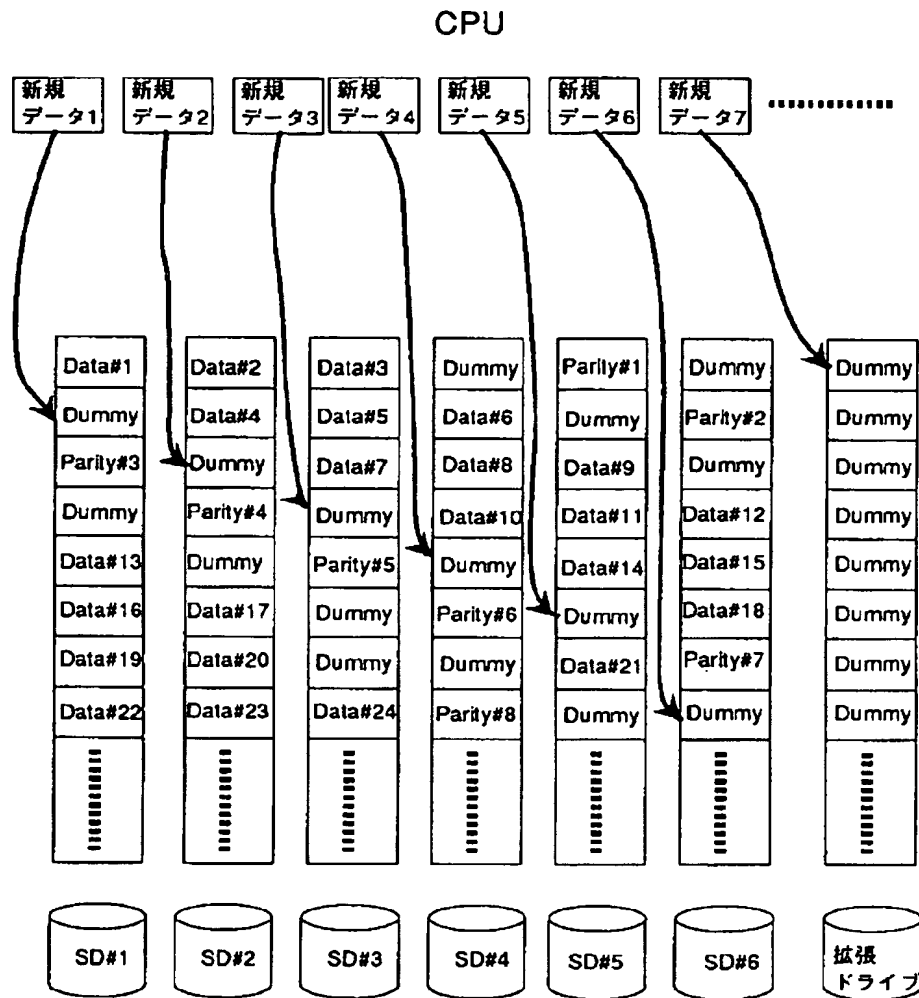
【図12】

図 12



【図13】

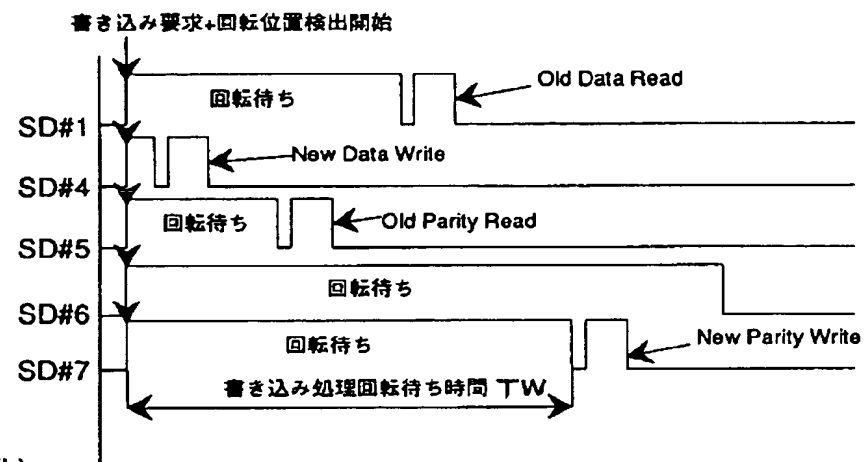
図 13



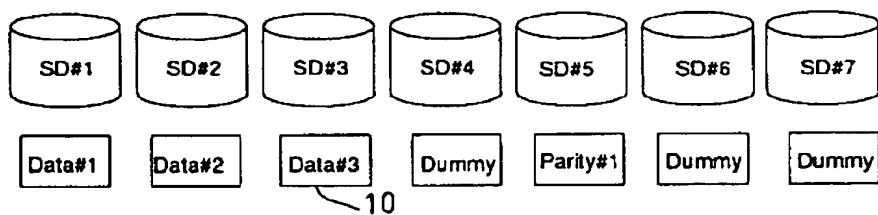
【図 1 4】

図 14

(a)

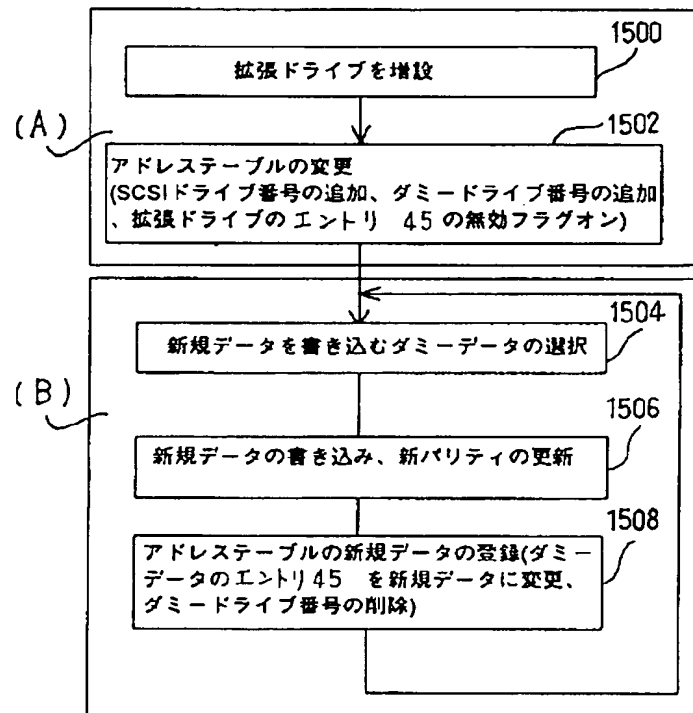


(b)



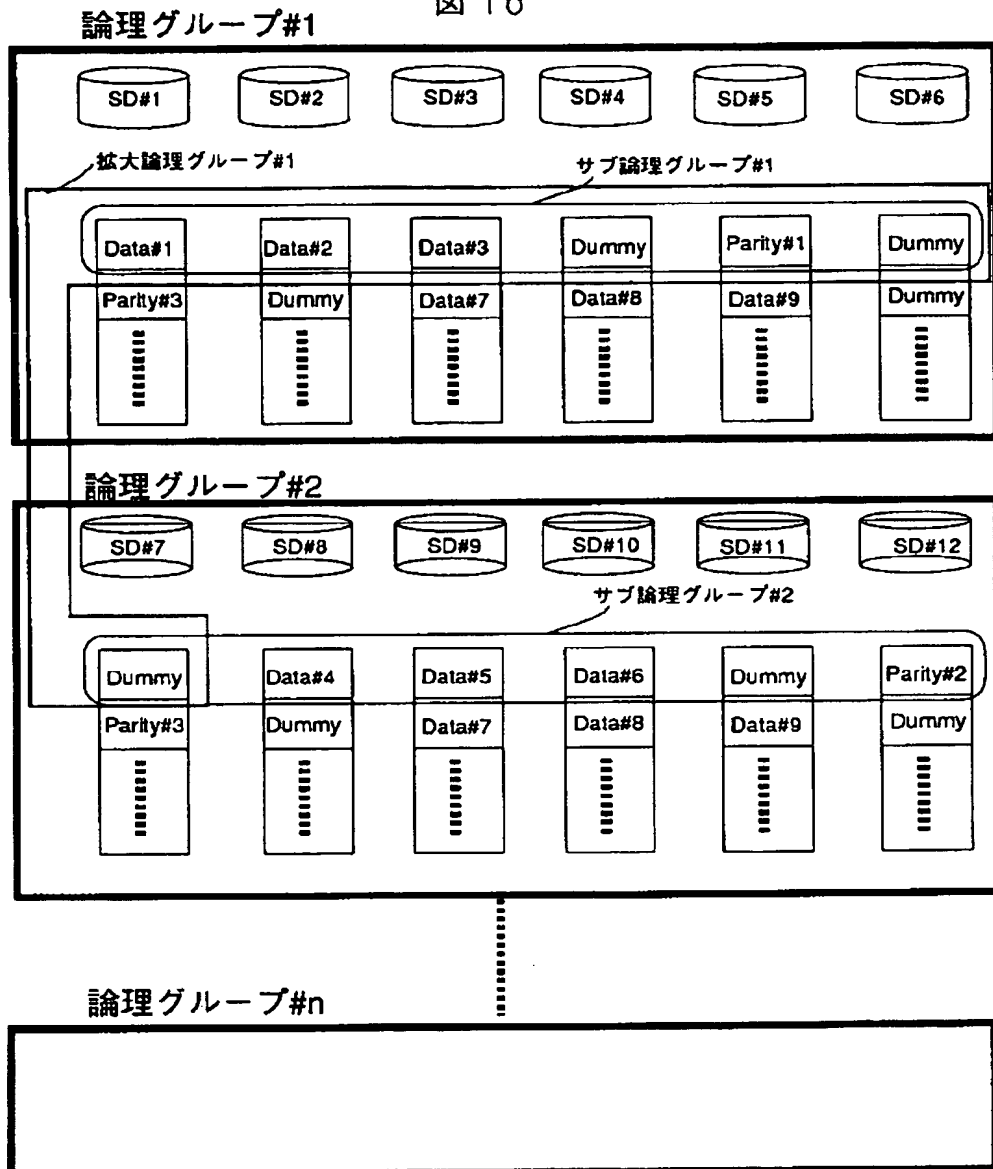
【図 1 5】

図 15



【図16】

図 16



【図21】

図 21

